Statistical Methods of Detecting Vertebral Fractures

Thesis

Statistical Methods of Detecting Vertebral Fractures

A thesis submitted for the degree of Doctor of Philosophy

by

Mark Lunt B.Sc, M.Sc

Department of Statistics

Faculty of Mathematics and Computing

January 2003

Submission date: 24 January 2003
Award date: 18 July 2003

Abstract

This thesis is concerned with the identification of vertebral fractures from measurements made of the anterior, mid and posterior heights of individual vertebrae. Two distinct problems are addressed: identifying deformities that exist at a particular moment in time from a single radiograph (prevalent deformities); and identifying deformities that have occurred between two consecutive radiographs (incident deformities).

A number of different statistical models for the vertebral heights are proposed, and compared to two existing methods in common use. The new models proposed are:

1. A number of polynomial models
2. A factor analysis model
3. An imputation based regression model

The polynomial models were fitted using both least squares and a robust method. In the simplest polynomial model, a single magnification factor was fitted for each subject, allowing for variation in size of the spine, but not variation in shape. More complex models in which the magnification factor was allowed to vary within an individual were also used. In addition, an outlier detection method is also applied to the data to detect subjects with fractures, and this method is also compared to the existing methods.

Models were compared not only on how well they predicted vertebral heights,

but also on how well they can identify fractured vertebrae and identify individuals with fractures.

Two approaches to identifying incident fractures are presented:

- Identify vertebrae that are classed as prevalent fractures on the second radiograph but not on the first radiograph;
- Identify vertebrae in which at least one height has shown a substantial reduction in height between the two radiographs.

It is shown that combining these two approaches has advantages over using either approach individually.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank David Hand, my supervisor, not only for the practical help and advice he gave me, but also for reviving my flagging enthusiasm when necessary.

I must also thank many colleagues: at the Institute of Public Health in Cambridge, where I started this work; at the Arthritis Research Campaign Epidemiology Unit in Manchester, where I completed it; and all those involved in the EVOS study.

I would like to thank my parents, for teaching me that I could achieve anything I wanted if I set my mind to it, and my wife, Lies, for finally convincing me that they might actually be right. She also deserves my thanks for the patience with which she tolerated the many hours I spend locked away in the study, as do my children Marie and Emma for understanding when Daddy had to sing their lullabies in a hurry.

Publications Arising From The Current Studies

Papers

1. **Lunt M**, Gowin W, Johnell O, Armbrecht G, Felsenberg D & Reeve J. *A statistical method to minimise magnification errors in serial vertebral x-rays.* Osteoporosis International, **12**:909–913, 2001.

2. **Lunt M**, Ismail A, Felsenberg D, Cooper C, Kanis J, Reeve J et al. *Defining incident vertebral deformities in population studies: A comparison of morphometric criteria.* Osteoporosis International, **13**:809–815, 2002.

In addition, the principle of defining incident deformities using a combination of height loss and unusual shape definitions, proposed in Chapter 14, has been adopted by the EPOS study, and used in the following papers:

1. The European Prospective Osteoporosis (EPOS) Study Group. *Incidence of vertebral fracture in Europe: Results from the European Prospective Osteoporosis Study (EPOS).* Journal of Bone and Mineral Research, **17**(4):716–724, 2002.

2. Vergnaud P, **Lunt M**, Scheidt-Nave C, Poor G, Gennari C, Hoszowski K et al. *Is the predictive power of previous fractures for new spine and non-spine fractures associated with biochemical evidence of altered bone remodelling ? The EPOS study.* Clinica Chimica Acta, **322**:121–132, 2002.

3. O'Neill TW, **Lunt M**, Silman AJ, Felsenberg D, Benevolenskaya L, Bhalla A et al. *The relationship between bone density and incident vertebral fracture in*

men and women. Journal of Bone and Mineral Research, **17**(12):2214–2221, 2002.

4. Roy DK, O'Neill TW, Finn JD, **Lunt M**, Silman AJ, Felsenberg D et al. *Determinants of incident vertebral fracture in men and women : Results from the European Prospective Osteoporosis Study (EPOS).* Osteoporosis International, **14**:19–26, 2003.

Part I.

Introduction

1. Vertebral Fractures

1.1. Background

Vertebral fractures are a common problem in osteoporosis, and are widely used as an endpoint in both epidemiological studies and clinical trials. Initially, a vertebral fracture was diagnosed by a radiologist reading an x-ray and giving a clinical opinion. However, this method was criticised as being too subjective (particularly for clinical trials). It has since been shown that both between-observer and within-observer variability is considerable.

It was thought that making measurements on the x-ray and and basing the diagnosis on these measurements would be more objective. A vertebral fracture will result in a loss of height in the vertebral body, so measuring such height loss should make it possible to identify fractured vertebrae. There are two possible situations:

A Prevalent Fracture: A single x-ray is available, and a fracture has to be diagnosed by the fact that it is an unusual shape, or the fact that one or more heights

are less than would be expected given the heights of adjacent vertebrae.

An Incident Fracture: Two or more x-rays are available, and a fracture has occurred in the interval between films. Such fractures may be identified in the same way as prevalent fractures outlined above. However, there is now additional information available from the additional films: incident fractures may be identified by the amount of change between two films.

Methods of identifying vertebral fractures based on measurements of vertebral bodies are referred to as morphometric methods. A number of such methods have been proposed, for identifying both prevalent and incident fractures. However, they can be re-expressed as models that predict individual heights, together with rules to determine whether a given height is unusually low based on the residuals from the model.

1.2. Problems to be addressed

Given these existing models as a baseline, we aimed to develop alternative models. Two families of models were used, polynomial and latent variable models. In the polynomial models, the vertebral heights were predicted as a polynomial function of the vertebral level, with a magnification factor for each subject. Thus the assumption is that all spines are the same shape, and differ from each other only in size. The latent variable models can allow for variation in shape as well as size. Both types of models were developed to use the data from either a single x-ray, to detect prevalent fractures, and from two consecutive x-rays to detect incident fractures.

1.2.1. Robustness

There are three major problems in predicting vertebral heights from measurements of other vertebral heights. The measured heights may be less than they should be due to fracture. These unreliable heights can cause two of these three problems with predicting heights.

Firstly, they can affect the model parameters. If the parameters of our model are biased, then it will be impossible to obtain accurate predictions of the vertebral heights. A method will be described as capable of *robust model definition* if it is able to produce unbiased estimates for the model parameters even when some of the heights in the sample in which the model is being defined are reduced by fractures.

Secondly, if unreliable heights are used in the model as predictors, then the prediction will again be unreliable, even if the model parameters are unbiased. Ideally, we would like to develop a method that can produce reliable predictions even if some measured heights are unreliable. A model that can do this will be described as capable of *robust model fitting.*

Initially, both of the above models were developed in subjects deemed to be free from fractures by an experienced radiologist. However, they should also be applicable to populations which may contain fractures. They therefore need to be robust, both in model definition and model fitting. That is, the regularities in shape identified in the model definition process should not be affected by the existence of deformities in the definition sample. Also, the heights predicted from the model should be unaffected by the presence of one or more deformities in the spine.

The third major problem is that of missing data, which are a very common problem in vertebral morphometry. It can be difficult to visualise a vertebra on a radiograph, and if it cannot be seen clearly, it cannot be measured. Therefore, any method developed must be applicable to subjects with missing data. Models that are capable of producing height estimates even when some of the heights have not been measured will be described as *robust to missing data.* A good morphometric model will have all three of these robustness properties.

2. Literature Review

2.1. Introduction

Since the early 1960's, there has been concern that the clinical identification of vertebral fractures from radiographs was too subjective, and agreement between radiologists was not sufficiently strong. Two general approaches have been taken to this problem. One is to attempt to improve agreement using protocols of fracture definitions, atlases of standard films and grading the severity of the fracture using predefined categories. This approach is referred to as qualitative or semi-quantitative.

The alternative is to replace the radiologist with an objective calculation based on measurements made on the x-ray film. Digital Vertebral Morphometry is the name given to a variety of such techniques to identify deformed vertebrae. All the techniques start with an image of the spine, on which one or more indices are measured. A threshold for each index is defined, and if a given index lies beyond

its threshold, that vertebra or spine, depending on the particular technique used, is classed as deformed.

The aim of Digital Vertebral Morphometry is to detect fractured vertebrae. However, there are conditions other than fracture that can lead to changes in shape in a vertebra and it thus being classed as deformed. It is therefore customary to use the word 'deformity' to describe a vertebra that satisfies one of the algorithms below, and reserve the word 'fracture' for a clinical reading of an X-Ray.

2.2. Radiology

2.2.1. Introduction

The spine contains 24 true vertebrae: the upper 7 are the cervical vertebrae, then there are 12 thoracic vertebrae, and labelled T1 (the highest) to T12 (the lowest). Below the thoracic vertebrae are the lumbar vertebrae, labelled L1 - L5. Most investigators are concerned only with deformities of the vertebrae from T4 to L4 or L5.

The main weight-bearing part of each vertebra consists of a ring of dense (cortical) bone, filled with porous (trabecular) bone. The top and bottom of the vertebra consist of plates of cortical bone, called the endplates. There are other bony structures at the posterior part of the vertebra, but they are not part of the vertebral body which is measured.

2.2.2. Conventional X-Rays

A conventional X-ray film is made from a point source of x-rays, which then pass through the subject and form an image on a film. For vertebral morphometry, a lateral image is needed, which is usually taken with the subject lying on their side, the source above them and the film beneath them. The exposure takes several seconds, during which time the subject is encouraged to breathe normally. This causes the lungs and ribs to move, blurring them on the image and making the vertebrae easier to see. Figure 2.1 illustrates the method.

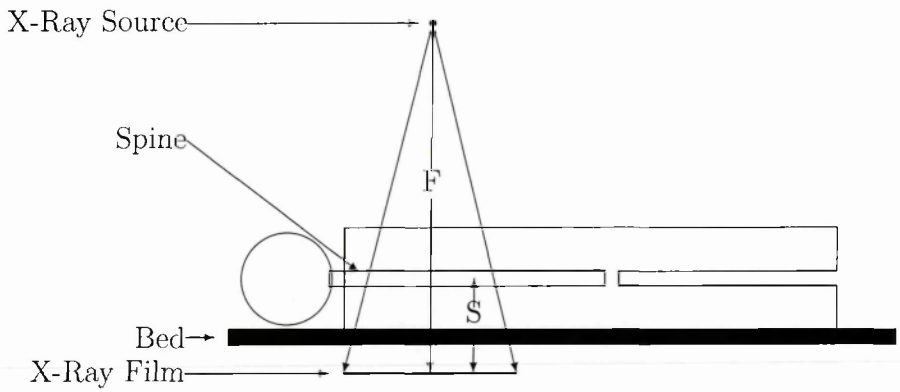


Figure 2.1.: A Subject About to be X-Rayed

If the film-focus distance (F)and the spine-film distance(S) are both known, then the magnification of the image can be calculated ($m = \frac{F}{F-S}$). The film focus distance is generally fixed, but the spine film distance depends on the subject being measured, increasing with the size of the subject.

It is very important that the spine is perfectly straight and horizontal. If this is not the case, the ratio of the focus-spine distance to the spine-film distance will

differ along the spine, and hence the magnifications of different vertebrae will differ, changing the apparent shape of the spine.

Another potential problem illustrated above is the fact that the x-rays are not perpendicular to the film at its ends, although they are in the centre of the film. This leads to the image of the vertebrae at the ends of the film appearing to be rotated.

Identifying the vertebrae from the image may be difficult, particularly in osteoporotic patients in whom the x-ray image may be faint. In such cases, the radiologist may identify a particular vertebra before making the scan, and place a marker (a small lead disc) on the bed at the level of the identified vertebra. This shows up clearly on the x-ray and the vertebrae are then easy to identify.

Figure 2.2 shows how an image of a vertebra on a radiograph is marked in order to measure the vertebral heights.

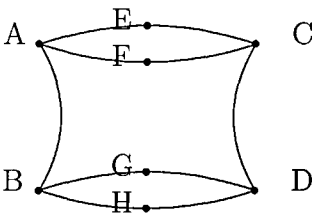


Figure 2.2.: A Vertebra Marked For Measuring

The distance AB is the anterior height Ha_i, and the distance CD is the posterior height Hp_i, where the suffix i indicates the vertebra on which the height was measured. The mid height is more difficult to define, since the endplates (AECF

and BGDH) can give rise to

1. a single line, if the images of the left and right rims of the endplate coincide.

2. two distinct lines, representing the left and right rims. This commonly occurs for vertebrae towards the edge of the x-ray, since the point source of x-rays makes such vertebrae appear to be rotated.

3. three distinct lines, representing the left and right rims and an image of the central cortex of the endplate. This only occurs if the vertebra appears to be rotated, and the cortex is sufficiently dense and caught at just the right angle to attenuate the x-rays strongly.

In case 1, the points may simply be placed on the lines. If two lines are visible, as in the diagram above, the points E,F,G and H are all marked. Then the left mid height is given by the distance EG, and the right mid height by the distance FH. The mid height, $Hm_i = \frac{EG + FH}{2}$. Hurxthal [1] recommends using the line representing the cortex, if it can be determined which is the cortex.

A further problem with determining the mid height is a benign condition known as Schmorl's nodes. These appear on an x-ray as a second line at the lower edge of the vertebra, but curving into the vertebral body. They can be easily identified in a good quality x-ray, but in a poorer quality image, typical in osteoporotic patients, they may be misleading as the image of the true endplate may be faint [2].

It is not possible to produce an image of all vertebrae from T4 to L4 on a single film, so two separate films must be used. The magnification of the two films may

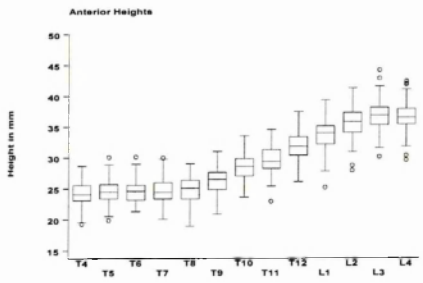
differ slightly, which will alter the apparent shape of the spine, if this is not corrected for.

The heights of the vertebrae in the lower spine tend to be larger than the heights in the upper spine. Also, the mid-height is generally slightly less than the anterior or posterior heights. The anterior height is generally slightly less than the posterior height in normal vertebrae, although this may not be the case for L4, which is often slightly wedged in the opposite direction. Figure 2.3 shows the distribution of heights in 126 clinically normal men and 134 normal women (taken from the Heidelberg EPOS centre).

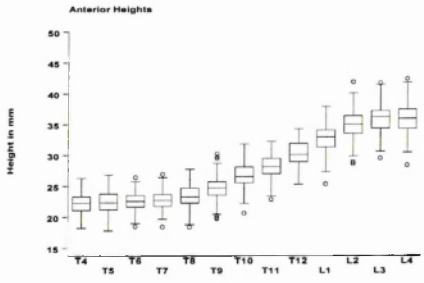
2.2.3. Morphometric X-Ray Absorptiometry (MXA)

MXA is fairly new technology, based on Dual Energy X-Ray absorptiometry (DXA), a technique widely used to measure bone mineral density. In DXA, a very low dose of X-radiation is passed through the subject and its attenuation used to measure the amount of bone and soft-tissue it has passed through. By using high resolution X-Ray detectors, it is possible to produce an image of the spine at the same time as measuring its density.

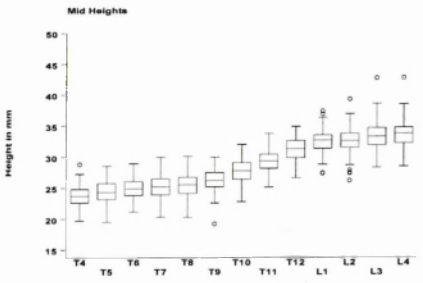
MXA does not produce images of the same quality as conventional radiographs. Most researchers agree that reproducibility of measurements on an individual image are less good [3]. However, MXA does not suffer from the magnification or rotation effects associated with conventional x-rays, and hence repeat measurements on individual subjects are likely to be more similar. Thus, MXA is likely to be less good for detecting prevalent deformities, but may perform as well as or better than x-rays



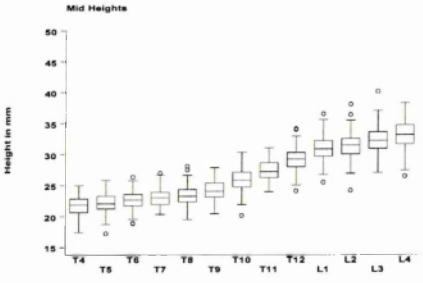
(a) Anterior Heights in Men



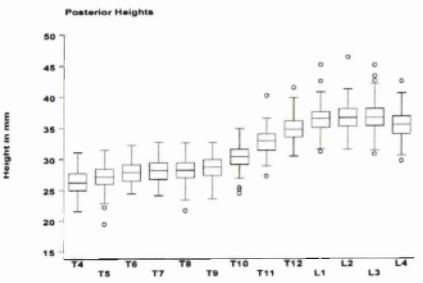
(b) Anterior Heights in Women



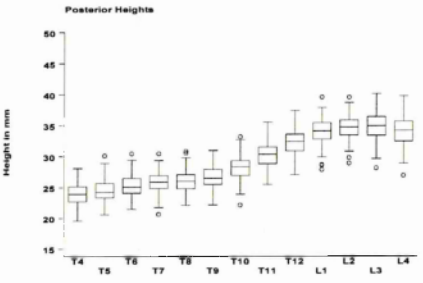
(c) Mid Heights in Men



(d) Mid Heights in Women



(e) Posterior Heights in Men



(f) Posterior Heights in Women

Figure 2.3.: Vertebral Heights in Normal Subjects from Heidelberg

in detecting incident deformities.

A further advantage of MXA is in automatic point placement. Edge detection algorithms are already built into MXA machines to enable them to calculate bone mineral density. Thus the points needed to define the vertebrae may be placed without the need for a human operator, which removes a possible source of measurement error. However, a number of groups are working on methods to determine the edges of vertebrae from digitised images of x-rays, so this advantage of MXA may be short-lived.

2.3. Detecting Prevalent Deformities

Methods for detecting prevalent deformities assume that only a single image of the spine is available. They aim to identify vertebrae that are unusual, compared to a reference population, based on this single image. An index of deformity is defined, which may be based on any measurements made on the image (height, area, shape ...).

Having chosen an index of vertebral deformity, it is then necessary to choose a range for that index that is to be considered 'normal' in order to be able to classify vertebrae as deformed. This normal range may be given either as an absolute range (e.g. if the ratio of the anterior height to the posterior height is less than 0.85), or based on the mean and standard deviation observed in the sample. However, since there may be fractures in the sample, robust methods are often used to calculate the mean and SD. These will be outlined later, since they seem to be particular to this

area. Some authors use radiographs that have been declared normal by an expert radiologist to define their normal range, but this can be criticised as retaining the subjectivity that digital morphometry was meant to replace.

Since a fracture results in a reduction of the vertebral height, the measured heights themselves may be used [4]. This method has been widely criticised, however, since taller subjects have larger vertebrae, and are therefore less likely to be classed as having a deformity for a given reduction in height. Current methods have some way of correcting for size, and define deformities as vertebrae with an unusual shape, although vertebral area has also been used [5].

2.3.1. Early Algorithms

The earliest published reference to the use of measurements on a radiograph to detect osteoporosis is by Barnett and Nordin[2]. They measured the mid and anterior heights of a single vertebra (they chose the vertebra with the best image, usually L3), and took the ratio $\frac{Hm}{Ha}$. Multiplying this by 100 gave what they termed the Spine Score, and they chose an arbitrary threshold of 80 (based on radiographs of 150 normal subjects), and a spine score below this threshold was considered to determine the presence of osteoporosis. It is interesting to note that they viewed the process whereby a vertebra changed shape as a gradual, continuing one. Currently, most methods dichotomise vertebra as either normal or deformed (with the exception of Minne's [6]), and are concerned with whether the vertebra has suffered a catastrophic change in shape, not a gradual one.

One of the earliest morphometric definitions of vertebral deformity that was

concerned with individual vertebrae was given by Gallagher et al. [7]. They gave 5 criteria for deformity: if a vertebra satisfied any one of them, it was classed as deformed. The 5 criteria were

1. H_a/H_p is outside the normal range (mean $\pm$ 2 SD). This measure was referred to as the "wedge angle", since is gives a measure of H_a as a proportion of H_p i.e. how far from parallel are the endplates of the vertebra. However, without knowing the depth of the vertebra (BD in figure 2.2), it is not possible to calculate the angle between the endplates.

2. The percentage reduction in anterior compared to posterior height is outside the normal range (mean $-$ 2SD).

3. The percentage difference in anterior height between adjoining vertebrae is outside the normal range (mean $\pm$ 2SD)

4. The anterior height is below the normal range (mean $-$ 2 SD)

5. The surface area of the vertebra on the X-ray is below the normal range (mean $-$ 2 SD)

One major problem with the above definition is that of multiple testing. If 5 tests are applied to each vertebra, and 13 vertebrae for each subject are tested (T4 to L4), 65 tests per subject will be performed. Tests 2, 4 and 5 are one-sided and would therefore have a type I error rate of 2.5%, whilst the error rate for tests 1 and 3 would be 5%, since they are two-sided. If the tests were independent, that would lead to an overall error-rate of 16% per vertebra and an error rate of over 90% per

subject (i.e. we would expect 90% of normal subjects to be classed as having at least one deformed vertebra). In practice, there are strong correlations between the different tests, and the false positive rate is far lower, but it is still a concern.

2.3.2. Vertebral Ratio Methods

Melton looked at 4 ratio measures per vertebra:

$$\begin{aligned} ap_i &= \frac{Ha_i}{Hp_i} \\ mp_i &= \frac{Hm_i}{Hp_i} \\ ppup_i &= \frac{Hp_i}{Hp_{i-1}} \\ ppdn_i &= \frac{Hp_i}{Hp_{i+1}} \end{aligned}$$

Initially, a threshold of 0.85 was chosen for all of these ratios, for every vertebra [8]. However, it was found that this method still led to a large number of false positives, since it did not take into account the fact that vertebral heights increase as one moves down the spine (and hence the mean of $ppdn_i$ was less than 1 for most i), nor that the natural shape of the vertebrae differed at different levels in the spine (Ha being naturally less than Hp for the thoracic vertebrae, but not for the lumbar vertebrae). Therefore an adjustment to each ratio was introduced to account for the mean value in the population differing from 1 for that particular ratio at that particular vertebral level. The adjustments were first calculated from a sample of 52 women without clinically evident fractures. A threshold of 0.85 was retained, as a change of about 15% was considered detectable by a radiologist reading an

x-ray. However, since the ap, mp and ppup ratios were generally less than 1, this adjustment had the effect of reducing the number of vertebrae classed as deformed.

In a comparison of the adjusted and unadjusted algorithms in 200 women aged over 50, Melton et al found that the unadjusted algorithm classed 83% of the women as having at least one deformity, whilst the adjusted algorithm only classed 26% as having a deformity. An experienced clinical reader diagnosed 28% of the women as having at least one deformity, and agreed with the adjusted algorithm for 90% of the women [8].

Deformities were classed into 3 categories. If ppup or ppdn was less than 0.85, the vertebra was classed as a crush deformity. If not, then if ap was less than 0.85, the vertebra was classed as a wedge deformity. Vertebrae that were not wedges but had mp < 0.85 were classed as biconcavities.

Eastell [9] suggested using the mean and standard deviation, rather than a fixed value, to define the threshold. The intention was that the specificity of the method was then fixed. He compared a number of possible choices of threshold to define deformities. He calculated the mean and standard deviation of each of the ratios in a sample of 42 women without radiological evidence of fracture, and compared thresholds of 2, 2.5, 3, 3.5 and 4 standard deviations to thresholds of 15% and 20%. He defined two types of deformity: a mild deformity in which one of the ratios was between 3 and 4 standard deviations below the population mean, and a severe deformity in which one or more of the ratios was more than 4 standard deviations below the mean.

A major problem with using the observed standard deviation to define the thre-

shold is that the observed variance comprises both natural variation in the population and measurement error. Therefore, two investigators measuring the same sample may have different thresholds, depending on how accurately they were able to make their measurements. It is possible for a fracture to change the shape of a vertebra sufficiently to be classed as a deformity by one investigator, but not the other: i.e. the sensitivity (the proportion of true fractures that are classified correctly) of the method decreases as the measurement error increases. Comparing prevalences between studies is therefore problematical.

Using a fixed threshold means that the sensitivity of the method is independent of the measurement precision. However, the specificity (the proportion of non-fractures that are classified correctly) now depends on the precision: poor precision means normal heights are more likely to appear to be reduced. Thus, comparing prevalences between studies is still a problem.

A second problem of the standard deviation approach is that a deformity is defined in terms of the mean and standard deviation measured in a population that does not contain any deformities: there is a circular argument there. If morphometry was to be self-contained, it needed to develop a way to calculate the mean and standard deviation even if the population used did contain deformities. In other words, robust estimators of the mean and standard deviation needed to be used. The methods used are outlined below in 2.5.2

One further refinement to this general method was proposed by Black[10]. Two comparisons are made for each posterior height, one to that of the vertebra above and one to that of the vertebra below. The fact that a vertebra is considered a

deformity if *either one* of these comparisons show the height to be reduced means that false positives are more likely to occur than at the anterior or mid heights (where only a single comparison is made). Furthermore, posterior deformities are the least common, since the vertebra is strongest at that point. He argued that if there had been a genuine reduction in height at the posterior, the vertebra would be completely collapsed, so there should also have been a reduction in anterior height. He therefore did not class a vertebra as deformed unless

$$Hp_i/Hp_{i-1} < Cp_{i1} \quad and \; Ha_i/Ha_{i-1} < Ca_{i1}$$

$$Hp_i/Hp_{i+1} < Cp_{i2} \quad and \; Ha_i/Ha_{i+1} < Ca_{i2}$$

where $Cp_{i.}$ are the appropriate level-specific cut-points.

The main criticism that can be leveled at all of the above methods is that they are inefficient. In each ratio, there is considerable measurement error in both the numerator and the denominator, and hence the precision with which the ratio can be measured is poor. In fact, using ratios in this way is mathematically equivalent to using regression to predict each height from a single other measured height (strictly, a weighted regression constrained to pass through the origin, see Section 4.2.1). Since there are great similarities in shape between different spines, it is reasonable to assume that it is possible to predict heights more accurately by using the heights of several adjacent vertebrae as predictors (multiple regression being an obvious candidate). Both Ross and Minne have devised methods that use a slightly different approach to predicting heights, but neither is completely successful.

2.3.3. Within Subject Comparison

Ross was one of the workers who originally used measured heights to define deformities. For instance, in [4], prevalent deformities were defined as vertebrae in which the anterior or posterior heights were more than 3 standard deviations below the vertebra-specific population mean. However, this method has been criticised since it makes false positives more common in shorter subjects and true positives less common in taller ones.

To get round this problem, Ross suggested a fairly complex procedure. First, all vertebral heights are measured and vertebra specific means and standard deviations of the heights calculated. Then each measurement is converted to a z-score. This is a way to allow for subjects being different sizes, since a tall subject will tend to have all positive z-scores (i.e. all vertebrae are larger than the population mean), whilst a short subject will have all negative z-scores. The mean(Z_j) and standard deviation (ZSD_j) (where j represents the measurement site (anterior, mid or posterior)) of these z-scores were then calculated over all vertebrae for each subject, after excluding any z-scores less than -3 or greater than 3. Then the population mean for ZSD_j, $PZSD_j$, was used as a measure of how far vertebrae would typically vary from the individual subject's mean z-score Z_j. Any vertebra in which the z-score was less than $Z_j - 3.0 \times PZSD_j$ was classed as a deformity, by analogy with the 3 standard deviations below the mean definition used elsewhere.

2.3.4. Overall Deformity – The Spinal Deformity Index SDI

All of the above methods attempt to classify vertebrae as either normal or deformed, and to sub-classify the different types of deformity. A different approach was proposed by Minne [6], namely to assess the degree of vertebral deformation in the spine. This approach may be of particular value in clinical trials, since a reduction in the degree of deformation due to the therapeutic agent may be detectable before a reduction in the fracture rate is. Minne's idea was to build a model of the heights in a normal spine, and measure how far from the predicted heights the measured heights were.

Rather than looking at ratios of heights within vertebrae or with adjacent vertebrae, they defined the height of the 4th thoracic vertebra to be 1, and divided all of the other heights by the relevant (i.e. anterior, posterior or mid) height of T4. They then plotted these three normalized heights against vertebral level, and fitted cubic regression equations to them. To select a threshold for each height, 110 subjects with radiologically normal spines were selected (73 women and 37 men), and the minimum value of each normalised height in this group identified. These minimum heights were smoothed by fitting a cubic equation to them, and any height lying below the predicted height from this cubic equation was classed as deformed.

To assess the degree of deformation of the spine, for any height that was below the threshold, the difference between the measured height and the threshold was recorded. The sum of all of these differences was referred to as the spine deformity index (SDI).

There are a number of problems with this method:

1. Only one of the measured heights is used to fit the model, which is inefficient.

2. It depends crucially on the height of T4: if there is a fracture or other deformity at T4, the method cannot be used. Even if T4 is not deformed, it is usually the uppermost vertebra on the image of the thoracic spine. This means that it is most susceptible to distortion due to the spine not being perfectly straight.

3. If T4 cannot be measured for some reason, the method cannot be applied.

2.3.5. Further Refinements to the Ratio Methods

The method developed by McCloskey and Kanis is a development of the Melton Eastell approach [11]. However, since it is quite a radical development, widely used and possibly the best algorithm currently available, I will discuss it in some depth.

McCloskey was aware that the large number of tests used in the Eastell algorithm could lead to a considerable number of false positives. However, using a more stringent threshold to improve the specificity would reduce the sensitivity of the algorithm. He therefore sought to improve specificity by insisting that a vertebra should satisfy two conditions before being classed as deformed, rather than one.

To create this second condition, he introduced the idea of the predicted posterior height, *Hpp*, calculated from the posterior heights of adjacent vertebrae. First the mean posterior height μ_i at each vertebral level i, in the population being considered is calculated. Then the predicted posterior height of the ith vertebra of a given subject, is calculated from the measured height of the jth vertebra in that subject,

Hp_j, as

$$Hpp_{ij} = Hp_j \times \frac{\mu_i}{\mu_j}$$

To improve accuracy, McCloskey took the mean of 4 predictions from measurements of the posterior heights of the four nearest vertebrae to calculate the predicted posterior height.

Then the ratio of the posterior height to the predicted posterior height was taken, and level specific means and standard deviations calculated. (The mean of this ratio is identically 1 if no trimming of unusual vertebrae is performed). Hence there are now 3 reference ranges for each vertebral level: $\frac{Ha}{Hp}$, $\frac{Hm}{Hp}$ and $\frac{Hp}{Hpp}$, with the thresholds set at 3 SD below the mean at Ca, Cm and Cp respectively.

A vertebra was classed as being deformed if any *two* of the following conditions were satisfied :

$$\begin{aligned}
\frac{Ha}{Hp} &< Ca \\
\frac{Ha}{Hpp} &< Ca \\
\frac{Hm}{Hp} &< Cm \\
\frac{Hm}{Hpp} &< Cm \\
\frac{Hp}{Hpp} &< Cp
\end{aligned}$$

The vertebra was classed as a biconcavity, wedge or crush, depending on which of the ratios lay outside the normal range. In a biconcavity, only $\frac{Hm}{Hp}$ and $\frac{Hm}{Hpp}$ are reduced. In a wedge, either the anterior and mid heights($\frac{Ha}{Hp}$ and $\frac{Hm}{Hp}$), anterior height

alone ($\frac{Ha}{Hp}$ and $\frac{Ha}{Hpp}$) or the posterior and mid heights ($\frac{Hp}{Hpp}$ and $\frac{Hm}{Hpp}$) are reduced. If both the posterior and anterior heights are reduced ($\frac{Hp}{Hpp}$ and $\frac{Ha}{Hpp}$) the vertebra is classed as a crush. Normally in a crush, the mid height is also reduced: there is no medical condition that would lead to reduction of the anterior and posterior heights without a reduction in the mid height, so the small number of deformities in which this occurs are probably false positives, due to an over-estimate of the predicted posterior height [12].

The algorithm was first applied to the highest vertebra on the x-ray (usually T4). Then it moved down the spine, evaluating each vertebra in turn. Each predicted posterior height was recalculated at this point, with any vertebra classified as having a posterior deformity removed from the calculation. Additionally, the four predicted posterior heights are ranked in order of size $Hpp_1..Hpp_4$ where Hpp_4 is the largest. Then for i=1 to 3, if

$$\frac{Hpp_i}{Hpp_4} < Cp$$

Hpp_i is not used to calculate the mean predicted posterior height. The intention of this procedure is to exclude from the calculations vertebrae that are deformed.

However, this may lead to bias in the mean predicted posterior height. Simulations suggest that if a sample of size 4 is drawn from a normal distribution (it is usually assumed that the vertebral height ratios follow a normal distribution in a fracture-free population), the probability that the difference in size between the largest and the smallest is greater than 3 standard deviations is close to 15%. Thus, in 15% of normal subjects, the smallest ratio will be below Cp and thus the smallest height will be removed from the calculations of Hpp. Thus, the mean will be biased

upwards. This may be a serious problem, since an over-large mean predicted posterior height is a sufficient condition for a vertebral deformity (if $\frac{Hp}{Hpp} < Cp$, Ha and Hm are only compared to Hpp, and thus are likely to also be considered reduced).

Taking the mean of 4 predicted heights is not equivalent to conventional multiple regression for predicting the posterior height. This has been done to facilitate removing a single predictor variable if the vertebra is found to be deformed. However, this method ignores any correlations that there may be between the predictor variables (which will be strong, around 0.8, for these vertebrae). Conventional multiple regression would therefore give more precise estimates. Furthermore, if one predictor variable needs to be removed from the regression equation, because the height has been reduced by a fracture or it could not be measured, it is still possible to obtain a predicted height using multiple regression, as will be shown in Chapter 10.

2.4. Incident Deformities

There are two basic approaches to defining an incident deformity. One is to measure the change between two x-rays, and see if the change is greater than some predefined limit. The second is to simply look for deformities on a second round of x-rays, using one of the methods outlined above, and then go back to the first round x-rays to see if they were already deformities. If not, they are classed as incident deformities. Those that were not prevalent are considered incident. McCloskey has recommended the second approach [11], although it does not allow for the possibility that a fractured

vertebra may deteriorate over time.

2.4.1. Change-based Methods

The limiting factor in the first approach is the measurement error. If the measurements made on each x-ray were extremely precise and accurate, changes would be easy to detect. However, errors in measurement are usually assumed to be normally distributed and independent, and neither of these assumptions hold true in this case. Slight changes in patient positioning can give rise to a change in magnification of the image, and thus the errors in measuring the three heights in a given vertebra are correlated. Furthermore, choosing exactly where to place the points on the image to measure can be problematic, and slight changes in the images can lead to large changes in the measured height. Thus the measurement errors follow a distribution that is close to normal around 0, but with overly large tails, and an apparently large reduction in height may be artifactual.

The first approach can have a variety of forms: changes in heights or changes in ratios, relative changes or absolute changes, fixed thresholds or thresholds based on measurement error. The arguments for and against the various types have been presented by NOF [13]. Currently, many authors consider that a change in height of at least 20%, provided that it is at least 4mm, should be considered an incident fracture[14].

There is a particular problem in determining whether a previously deformed vertebra has deteriorated in the time since the first x-ray. Measurements are more difficult to make on deformed vertebrae, so the variation in height due to measure-

ment error will tend to be larger. Furthermore, since the heights have already been reduced, a given absolute change in height represents a larger percentage change. There is thus a risk of over-diagnosis. For this reason, some researchers and trialists exclude such vertebrae from assessment, but this excludes any information thay may supply on how deformities progress. The alternative is to insist on a given change in height as well as a percentage change: the standard accepted by the Food and Drug Administration for clinical trials is 20% and 4mm.

2.4.2. Point-Prevalence Methods

In the second approach, the pairing of the films is ignored. The first round films are evaluated for deformities, and the second round films are evaluated completely independently. This method is not using all of the available information, and hence could be expected to be less efficient. In particular, the definition of a prevalent deformity relies on comparing a ratio to the variation of that ratio in a normal population, which will depend on both true variation between individuals and measurement imprecision. The change-based methods are only affected by measurement imprecision, not the variation between individuals. However, McCloskey *et al* have shown better specificity by defining incident deformities as vertebrae that were not classed as prevalent deformities on the first x-ray but were classed as prevalent deformities on the second [11]. They claimed this was due to the large tails in the distribution of measurement errors leading to a large number of false positives if changes in the vertebra are the only criterion used.

Another argument put forward to support this approach is that the definition

of a deformity remains the same, whether it is observed on the first occasion or the second. This is true, but it requires that we ignore the information gathered from the first x-ray when determining the status of the vertebra on the second occasion. Furthermore, a vertebra with a prevalent deformity whose shape is close to the threshold may be classed as normal on the first round and deformed on the second purely due to measurement error. This is not an incident fracture and would have to be considered a false positive.

It may be that a combination of the two approaches may be the best: insist that a vertebra is a prevalent deformity on the second round and that it has changed from the first round before classing it as an incident deformity. Some work that I did for the EVOS study investigating this idea is presented in Chapter 14, and has also been published [15].

2.5. Robust Estimation

2.5.1. Introduction

Statistical modelling is concerned with estimating certain parameters of a given population, such as the mean, standard deviation, regression coefficients etc. However, if some of the observations are not from the population of interest, the estimate of the parameter of interest may be quite different from its true value. For example, if we wish to calculate the mean vertebral height in unfractured vertebrae, but some of the heights in our sample have been reduced by fractures, the mean of our sample is likely to be less than the mean in the population of unfractured heights. Methods

of estimating parameters that are not unduly influenced by a small number of very unusual observations are known as "robust" methods.

For example, whilst the arithmetic mean of a sample can be affected greatly by a single very large (or very small) outlier, the sample median is much less affected. If the data are believed to be from a symmetric distribution, the expected value of the sample median is the same as the expected value of the sample mean, and so can be used as a robust estimate of the population mean.

2.5.2. Robust Estimation Methods Used in Vertebral Morphometry

The morphometric methods in current use require robust estimates of the mean and standard deviation of the various ratios of vertebral heights. The simplest method was an iterative trimming method presented by Melton et al[8]. For each ratio, the interquartile range was calculated, and any observations lying 1.5 interquartile ranges above the 75th percentile or below the 25th percentile were removed. This procedure was repeated until no further observations were removed, and the mean and standard deviation of the remaining observations used to provide the reference range.

Even with data drawn from a normal distribution, one might expect a small number of observations to be removed as outliers. This should not be a disadvantage for estimating the mean, since such observations are equally likely to be above the mean as below it. However, if the most extreme observations are removed, the standard deviation of the remaining sample must decrease, and hence this method will give an underestimate of the true population standard deviation.

A more sophisticated approach was suggested by Black et. al. [16]. They assumed that the distribution of the vertebral ratios was normal, but with an elongated tail due to deformities. After trimming the most extreme 5% of points from either end of the distribution, they estimated the mean of the distribution by first producing a histogram of the ratios, then plotting the natural log of the frequency in each interval against the mid-point of the interval. They then fitted a quadratic curve to these points, and took the highest point on the curve to be the population mean. They found that this procedure had little effect on the estimated mean, and also that the median was almost identical to the adjusted mean in every case.

They also showed that robust methods were necessary in estimating the standard deviation. This they did by drawing a Gaussian probability plot for the values, then fitting a line to the resulting plot after trimming 10% of the values from each end. This procedure led to a considerable reduction in the estimate of the standard deviation (around 10%) compared to the untrimmed standard deviation.

There is not, as far as I am aware, a direct comparison of this method of trimming to that of Melton. However, Rocke and Woodruff [17] have shown that Black's method is biased upwards in the presence of a moderate number of outliers. Melton's method, on the other hand, is biased downwards, particularly if the proportion of outliers is small.

3. Data Available

3.1. Introduction

The data used for this project were taken from the European Prospective Osteoporosis Study (EPOS). This study was undertaken to identify risk factors for vertebral fractures in a healthy population. All participants had a spinal radiograph taken on entry to the study, in order to identify subjects with vertebral fractures. Dietary, lifestyle and family history risk factors were recorded via a questionnaire. The prevalence of vertebral fracture could then be modelled as a function of the risk factors.

Men and women aged 50-80 years were recruited from population registers in 36 European centres. The samples were stratified into six 5-year agebands and by sex. The aim was to recruit 50 subjects from each stratum to give a total sample size in each centre of 600. After a period of approximately 4 years (the exact time varied between centres), subjects were recalled for a second x-ray.

3.2. Radiology

3.2.1. Measurements

In the first round of the EPOS study, 15,000 x-rays were taken in 36 European centres. A standard protocol was sent to each centre, so that all x-rays were produced in the same manner. Subjects were encouraged to breathe whilst the film was being taken, so that the soft tissue in the chest moved and became blurred on the film, but the spine remained still and was clearly visualised. The film-focus distance was fixed by the protocol at 120cm.

In order to enable the entire spine to be visualised, two films were required. One film was centred on T7 and the other on L1. Ideally, the spine-film difference should have been recorded for each film, to enable the measured heights to be corrected for magnification. However, there were two problems with this:

1. Some centres recorded only one spine-film distance (and some centres did not record any).

2. Some vertebrae (commonly T11 and T12) appeared on both films, and could be measured on either. There was no protocol for this: the radiologist measuring the film measured whichever film provided the better image. However, this means that we cannot be certain of the magnification of certain vertebra.

In addition, one centre provided x-rays already taken using a different protocol for an ongoing study, since they felt it would not be ethical to expose their population to an additional dose of radiation when films were available. However, these x-rays

consisted of only a single film, and so the entire region from T4 to L4 did not appear.

All x-rays were sent to Berlin to be measured. A single radiographer measured all x-rays, to eliminate inter-observer variation. The anterior, mid and posterior heights from T4 to L4 were measured and entered on a database, along with the spine film distance if available.

Only 6721 subjects in 29 centres took part in the second round of x-rays. The same protocol was used as for the first round. These were all measured a single radiographer (but not the radiographer who measured the first round x-rays). The anterior, mid and posterior heights were again entered into the database. Then the ratios of the anterior to posterior height and mid to posterior height was calculated for each vertebra on both occasions. If either ratio was below 0.75, or had changed by 0.15 or more, the senior radiologist reviewed the first and second round x-rays side-by-side, and made any adjustments to the placement of the points used to mark the vertebral heights that he thought necessary.

The drop-out rate in this study was considerable. In addition to the 7 centres who were unable to continue with the study, many subjects who had had an x-ray at baseline did not return for a second x-ray. A comparison of the baseline data of those who did return to that of those who did not revealed that the non-returners tended to be older, although the difference was small. They also differed in other risk factors for vertebral fracture, such as fracture history, but these differences could be explained by the difference in age [18].

How will this dropout affect the results of this thesis ? Firstly, the incidence of fracture will be lower. Secondly, the fractures in the subjects who did not return

may be smaller than those in the ones who did return, since older subjects tend to have larger fractures. However, as we will see, the problems arise in trying to identify smaller fractures: most methods can identify large fractures easily. We could therefore expect all methods to perform better in a true population sample, although there is no reason to suppose that the relative performance of the methods would change.

3.2.2. Clinical Readings

When the second round measurements were made, a clinical evaluation of each spine was made by the senior radiologist making the measurements. Each subject was given one of the following classifications, based on reading both films:

- No Deformity
- Osteoporotic fractures, no change between the two x-rays
- Osteoporotic fractures, at least one has either appeared or deteriorated since the first film
- Osteoarthritic damage
- Traumatic fracture
- Scheuermann's disease
- Congenital deformity

If the subject was classed as having one or more osteoporotic deformities, each vertebra was assessed individually. The shape of the vertebra on both the first round and second round x-rays was classed as one of the following

1. **Normal:** there was no apparent fracture in this vertebra
2. **Concave:** one of the two endplates of the vertebra was clearly fractured
3. **Biconcave:** both endplates of the vertebra were clearly fractured
4. **Crush:** the vertebral body had collapsed, causing a reduction in all three vertebral heights

In addition the total number of vertebral fractures in the spine on each film was recorded. However, it was not recorded exactly which vertebrae had incident fractures. In most cases it could be deduced, since the shape was either normal on the first round and not on the second, or at least had changed shape between the two rounds. However, there were some cases in which previously fractured vertebrae deformed further without changing from one type of fracture to another.

3.3. Description of the Populations

3.3.1. Subjects Used for Assessing Prevalent Deformities

Three centres were chosen to assess methods of defining prevalent deformities. It was necessary to use more than one centre, since it has been suggested that normal spines vary in shape between populations. The three centres which had the largest

numbers of x-rays on the second round were therefore chosen for this part of the analysis. The second round measurements were used because only approximately half of the subjects who had a baseline film returned for the follow-up film, and only subjects with both films could be included in the analysis of incident deformities: it seemed sensible to use the same subjects for analysing prevalent deformities.

The subjects in each centre were divided into four groups:

1. The training set
2. The testing set
3. Subjects with fractures
4. Subjects with deformities with other causes.

Subjects with no apparent deformities were divided into two samples, partly because fitting some of the models to large groups was problematic and partly to provide another sample for cross-validation of the model. The training sets consisted of 70 subjects of each sex randomly selected from all the subjects with no fractures and with no missing measurements. Any subjects with missing data were excluded from the training sample, because the Minne model cannot be fitted if the measurements of T4 are missing, whilst the other methods can. Thus, if missing values were permitted in the training sample, we would have different numbers of observations for the different methods, making them harder to compare. The testing set consisted of the remainder of the subjects with no fractures. The number of subjects in each of these groups is given in Table 3.1.

Population	Training	Testing	Fractures	Other Deformities	Total
Heidelberg Men	70	56	24	39	189
Heidelberg Women	70	64	20	12	166
Malmo Men	70	80	39	17	206
Malmo Women	70	96	40	9	215
Graz Men	70	83	35	19	207
Graz Women	70	110	20	9	209
Total	420	489	178	105	1192

Table 3.1.: Numbers of subjects in prevalent deformities analysis

In addition to the 178 subjects with fractures at the time of the second film, there were also 105 subjects who had deformities of other types: degenerative changes, congenital deformities, Scheuermann's disease etc.

As can be seen from table 3.1, 14.9 % of subjects had at least one vertebral fracture at the time of the second x-ray. A number of subjects had multiple fractures: the numbers of subjects are given in table 3.2.

Slightly more men than women have deformities (16.3% vs 13.6%), but the difference is not statistically significant ($p = 0.19$). However, the types of deformities did differ between men and women. The number of deformities of each type in men and women is given in table 3.3. It can be seen that only about 1/3 of deformities in men are concave or worse, compared to over 1/2 in women. The difference in distribution of shapes between men and women is statistically significant ($\chi^2(3) = 12.7422$, $p = 0.005$).

Number of Fractures	Number of Subjects (%)
0	1014 (85.1)
1	107 (9.0)
2	49 (4.1)
3	11 (0.9)
4	3 (0.3)
5	1 (0.1)
6	3 (0.3)
7	4 (0.3)

Table 3.2.: Distribution of numbers of prevalent fractures

Shape	Men	Women
Normal	7670	7525
Wedge	100	65
Concave	44	61
Biconcave	7	11
Crush	2	6
Unknown	3	2

Table 3.3.: Shapes of vertebral deformities in men and women

3.3.2. Subjects Used for Assessing Incident Deformities

Subjects from all 29 centres who took part in the second round of x-rays were used in assessing incident deformities. This was because incident deformities are less common than prevalent deformities: prevalent deformities include all deformities that have occured during an individuals lifetime, whilst incident deformities only include those which occured since the previous x-ray, a mean of 3.8 years earlier. So whilst there were over 300 prevalent deformities in the 1192 subjects from Heidelberg, Malmo and Graz, there were only 38 incident deformities in these subjects. By including all 6721 subjects from the 29 centres, a total of 295 fractures were obtained, which made comparisons more reasonable.

Part II.

Prevalent Fractures

4. Existing Models of the Spine

4.1. Introduction

In this chapter I will examine the three methods most commonly used currently to show how each method can be recast as a model to predict heights. I will also outline the steps that were taken to ensure that the model is both defined and fitted in a robust manner.

There were two reasons for regarding the existing methods as models. Firstly, one way to assess new models is to see how closely they can predict the heights of the vertebrae. It would therefore be useful to have predicted heights from the existing models for the sake of comparison.

Secondly, it has been argued that only ratios, not heights, should be used for identifying fractures, since ratios describe the shape of the vertebra, and it is by the shape of the vertebra that fractures are diagnosed clinically. The idea of comparing a height to its expected value was seen as completely different from how the existing

methods worked, and unsuitable for identifying fractures. Only by showing how existing methods could be thought of as comparing an observed height to an expected height could I convince my radiological colleagues that comparing observed heights to expected heights was a sensible thing to do.

4.2. The Melton-Eastell Algorithm

4.2.1. Predicting Heights

Although they do not use the word, the Melton-Eastell approach to defining vertebral deformities is based on a model of the spine. We have seen in section 2.3.2 that the Melton-Eastell algorithm is based on 4 ratios: $ap_i = \frac{Ha_i}{Hp_i}$, $mp_i = \frac{Hm_i}{Hp_i}$, $ppup_i = \frac{Hp_i}{Hp_{i-1}}$, and $ppdn_i = \frac{Hp_i}{Hp_{i+1}}$. They concern themselves with the *ratios* of vertebral heights, as measures of the shape of the vertebrae , rather than the heights themselves.

However, the assumption of the method is that each ratio is normally distributed with mean μ and variance σ^2. Consider, for example, the ratio $\frac{Ha}{Hp}$ in the j^{th} vertebra in the i^{th} subject. The Melton-Eastell approach assumes that this is normally distributed with mean μ_{apj} and variance σ^2_{apj} and assigns a z-score to the i^{th} subject of

$$\frac{\frac{Ha_{ij}}{Hp_{ij}} - \mu_{apj}}{\sigma_{apj}}$$

If, instead, we approach the problem as a regression equation to predict Ha_{ij} from Hp_{ij}, we get the regression equation

$$Ha_{ij} = a + bHp_{ij} + \epsilon_{ij}$$

If we constrain the regression to pass through the origin, and weight each observation by $\frac{1}{Hp_{ij}}$, we get

$$\begin{aligned} Ha_{ij} &= bHp_{ij} + Hp_{ij}\epsilon_{ij} \\ \Rightarrow \epsilon_{ij} &= \frac{Ha_{ij} - b\ Hp_{ij}}{Hp_{ij}} \\ &= \frac{Ha_{ij}}{Hp_{ij}} - b \end{aligned}$$

To obtain a least squares estimator for b, the above expression is squared, summed over all subjects, differentiated with respect to b, and set equal to 0. We then get

$$\begin{aligned} \sum_{i=1}^{n} \epsilon_{ij}^2 &= \sum_{i=1}^{n} \left(\frac{Ha_{ij}}{Hp_{ij}} - b \right)^2 \\ &= \sum_{i=1}^{n} \left(\frac{Ha_{ij}}{Hp_{ij}}^2 - 2b\frac{Ha_{ij}}{Hp_{ij}} + b^2 \right) \\ \Rightarrow \frac{\partial}{\partial b} \sum_{i=1}^{n} \epsilon_{ij}^2 &= \sum_{i=1}^{n} \left(-\frac{2Ha_{ij}}{Hp_{ij}} + 2b \right) \\ &= 0 \\ \Rightarrow 2nb &= \sum_{i=1}^{n} \frac{2Ha_{ij}}{Hp_{ij}} \\ \Rightarrow b &= \frac{\sum_{i=1}^{n} \frac{Ha_{ij}}{Hp_{ij}}}{n} \end{aligned}$$

which is exactly the definition of the mean of the ratio $\frac{Ha_i}{Hp_i}$.

If we wish to arrive at the same estimate for σ_{apj} using this regression approach as using Melton's original ratio approach, it is necessary to use the same trimming algorithm as is used by Melton. He applied it to the raw ratios, whereas in regression we would have to apply it to the residuals from the regression. However, in principle

it is possible to use a suitable form of robust regression to get exactly the same results as the Melton-Eastell algorithm.

Ha_{ij} may therefore be estimated, using this model, as the $Hp_{ij} \times \mu_{apj}$. Likewise the mid height may be estimated as the product of the individual's posterior height and the population mean of the mp ratio. For most vertebrae, two estimates of the posterior height are available: $\hat{ppup} \times Hp_{i(j-1)}$ and $\hat{ppdn} \times Hp_{i(j+1)}$. Since the vertebra will be classed as deformed if the observed height is significantly less than either of these expected values, the minimum of the two estimates of the posterior height can be used as the predicted value.

4.2.2. Robustness Concerns with Model Definition

The model definition consists of estimating the parameters μ and σ^2 for each ratio. If fractures are included in the sample used to define μ and σ^2, μ will be underestimated and σ overestimated. This will make fractures more difficult to detect. Melton was aware of this problem and used an iterative trimming algorithm described in section 2.5.2 to remove outlying ratios before calculating the mean and standard deviation.

This method should produce unbiased estimates of μ. However, as we have seen, trimming extreme observations will underestimate the variance unless steps are taken to avoid it.

4.2.3. Robustness Concerns with Model Fitting

No effort was made in this algorithm to provide robust model fitting. If a posterior height has been reduced by a fracture, any heights predicted from it will be smaller

than they should be. However, if either the vertebra above or the vertebra below the fractured vertebra is unaffected, the fractured height should be detected. Since the aim of the algorithm is only identifying fractures, not predicting heights, this lack of robustness is accepted. However, it should be noted that if there are three adjacent fractured vertebrae, there is no guarantee that the middle fracture will be detected as such.

4.2.4. Robustness to Missing Data

The anterior and mid heights are predicted from the posterior height of that vertebra. Generally, if it is possible to measure one height in a vertebra it is possible to measure all three. Therefore, there should be no problem predicting anterior and mid heights for those vertebrae that have been measured. However, the posterior height is predicted from the posterior heights of the adjacent vertebrae: if it was not possible to measure one of these, only one estimate of the posterior height is available. If neither adjacent vertebra could be measured, it is not possible to determine whether the posterior height was reduced.

4.3. McCloskey-Kanis

4.3.1. Predicting Heights

This method is slightly more difficult to recast as a model, since a ratio has to satisfy *two* criteria before being classed as a deformity. However, it is possible to obtain the expected values of each height. The predicted value of the posterior

height is obtained directly from the algorithm. Then two separate estimates can be made of the anterior and mid heights, one as the product of the measured posterior height and the mean anterior/posterior ratio or mid/posterior ratio for that vertebral level, and the other using the predicted posterior height and the appropriate ratio. If the posterior height is significantly lower than expected, the anterior/posterior and mid/posterior ratios are not used, so the predicted anterior and mid heights are given by the product of the mean anterior/posterior or mid/posterior ratio and the predicted posterior height. If the posterior height is not significantly less than expected, the algorithm will not declare the vertebra to be a fracture unless *both* of these heights are less than a fixed cut-off, so we can use the larger of the two heights as our predicted value.

4.3.2. Robustness Concerns with Model Definition

The model definition involves estimating the same parameters as in the Melton-Eastell model: the mean and standard deviation of a number of ratios. The same concerns with robustness apply, and the same solutions were adopted.

4.3.3. Robustness Concerns with Model Fitting

This method was devised in part to avoid the problem of the lack of robustness in fitting the Melton-Eastell algorithm when a posterior height is reduced by a fracture. The anterior and mid height are predicted using both the measured and a predicted posterior height. If the measured posterior height is much less than its predicted value, it is not used to predict the other heights. This provides robust estimates

of the anterior and mid heights in vertebrae with fractures affecting the posterior height that were not available with the Melton-Eastell method.

Even in the predicted posterior height, there is protection against using unusually low posterior heights. Four separate estimates of the posterior height are made and compared to each other. If one estimate is unusually low compared to the other three, it is excluded from the calculation of the predicted posterior height: the predicted posterior height is simply the mean of the estimates that have not been excluded. This provides added robustness to the predicted heights.

4.3.4. Robustness Concerns with Missing Data

Using several vertebra to predict the posterior height could have serious repercussions on the robustness to missing data. If conventional multivariate regression were used for prediction, no prediction would be available if *any* of the four adjacent vertebrae were unmeasured. However, because the prediction is made from each vertebra separately, and the mean of the four predicted values used as the predicted height, it is still possible to obtain an estimate unless *all* of the four adjacent vertebrae were unmeasured.

4.4. Minne

4.4.1. Predicting Heights

Although the word model is never used, there is an implicit model underlying the Minne definition of deformity. The assumption is that all spines are of a similar

shape, and differ only in size. Furthermore, it is assumed that the heights, after scaling by dividing by the corresponding height of T4, can be fitted to a cubic curve. Thus, the Minne model of the spine is

$$\frac{H_{sij}}{H_{s1j}} = a_s + b_s i + c_s i^2 + d_s i^3 + \epsilon_{sij}$$

where s is the site on the vertebra of the height (anterior, mid or posterior), i is the vertebral level (1 for T4, up to 13 for L4) and j identifies individual subjects. The predicted value of the height H_{sij} is then

$$H_{sij} = H_{s1j} \times \left(a_s + b_s i + c_s i^2 + d_s i^3\right) \tag{4.1}$$

4.4.2. Robustness Concerns with Model Definition

No attempt was made to deal with the problem of fractures in the sample used to define the model. It was explicitly stated that the model needed to be defined using a group of subjects free from vertebral fractures. That leaves a subjective element in the definition: since radiologists may disagree about the presence of a fracture, which radiologist should decide which subjects need to be excluded ?

4.4.3. Robustness Concerns with Model Fitting

Robust model fitting was a concern with this method. However, the authors do not appear to be aware of the statistical methods for robustly fitting models. Instead, they chose to use the three heights of a single vertebra (T4), and predict all other heights from those three. T4 was chosen because fractures are rare in this vertebra.

However, this is an inefficient way of fitting a model, since the measurements of the other 12 vertebrae are not used. It is also not robust: fractures may be rare in T4, but if there is one, all of the predicted heights will be too low, and other fractures in the spine may be missed.

4.4.4. Robustness Concerns with Missing Data

All predictions in this method are based on the heights of T4. Therefore, if T4 is not measured, the method cannot be used, although it is robust to missing data at any other vertebral level. However, since T4 is at the extreme of the x-ray film, it is more common for this vertebra to be unmeasureable than most other vertebrae.

5. Polynomial Models of the Spine with Fixed Magnification

5.1. Introduction

The simplest approach to modelling the heights of the vertebrae is to use linear regression. The heights tend to increase down the spine as was seen in figure 2.3. This increase is not, however, linear, so a polynomial will need to be fitted. It is apparent from figure 2.3 that a polynomial of at least order 3 will be required.

There are a number of questions that we would like these models to answer:

1. Are the natural shapes of the spine different in different centres ?
2. Are the natural shapes of the spine different between men and women ?
3. How accurately can we predict vertebral heights in normal subjects ?
4. How accurately can we identify vertebrae in which at least one height has been reduced by a vertebral fracture ?

5.2. Methods

5.2.1. Defining the Model

Initially, we assume that all spines of the same sex were the same shape, and differ only in size. Let the subscript s take the values 1, 2 and 3, for the anterior, mid and posterior heights respectively, and use the subscript i to represent the vertebral levels, with i=1 at T4 (the uppermost vertebra commonly measured) and i=13 at L4 (the lowest vertebra measured in our dataset). The subscript j represents the j^{th} subject. Then any model for vertebral heights H_{sij} that assumes spines are the same shape can be expressed as

$$H_{sij} = m_j \times f(s,i) + \epsilon_{sij}$$

where m_j can be thought of as a magnification factor, and $f(s,i)$ describes the shape of the spine. ϵ_{sij} represents the difference between the predicted height $m_j \times f(s,i)$ and the measured height H_{sij}. Minne actually fitted three separate magnification factors m_{sj}, one each for the anterior, mid and posterior heights, and thus allowed some variation between subjects in the shape of the spine. We will do the same, and see if it leads to an improvement in the fit of the model.

If we fit a polynomial of order r as the model, we get

$$f(s,i) = \sum_{k=0}^{k=r} a_{sk} \times i^k$$

The complete polynomial model

$$H_{sij} = m_{sj} \times \left(\sum_{k=0}^{k=r} a_{sk} \times i^k \right) + \epsilon_{sij} \tag{5.1}$$

is non-linear. It should also be pointed out that it is not identified: if all of the m_{sj} are multiplied by a constant c, and all the a_{sk} divided by the same constant, the predicted heights are not changed. A constraint needs to be placed on the model so that it is identified. The simplest approach would be to constrain the a_{sk} so that, for example, $\sum_k a_{sk} = 1$ for all s. However, since this is an arbitrary constraint, it makes the parameters difficult to interpret. An alternative constraint is to set the mean magnification factor over all individuals at each site, $\bar{m}_{s.}$, to be equal to 1. Using this constraint, the polynomial part of the model predicts the vertebral height of an 'average' individual, and the magnification factors give a measure of the overall height of an individual relative to this 'average' individual.

The model defined by 5.1 can be fitted straightforwardly using standard statistical software. The procedure `nl` in `stata` was used. Fitting a nonlinear model requires that sensible initial values are supplied for the parameters. This was done by assuming that the magnification factor was 1 for every subject. A cubic model was fitted to the data, and the parameters from this model used to initialise the a_{sk} parameters, with $a_{sk} = 0$ for $k > 3$.

Alternatively, the above model can be fitted as two separate linear models. First, we assume that $m_{sj} = 1$ for all sites and subjects, and fit a linear regression model to obtain initial estimates for the a_{sk} parameters. We can then obtain estimates for the m_{ij} parameters by fitting a linear regression model without an intercept. Having obtained estimates for the m_{ij}, we can calculate $z_{sij} = H_{sij}/m_{sj}$ and regress z_{sij} on $i^0, i^1, \ldots i^k$ (this time with an intercept term to give an estimate of a_{s0}) to obtain new estimates for the a_{sk} parameters. This regression needs to be weighted

by $1/m_{sj}^2$ to allow for the fact that the error term is now ϵ_{sij}/m_{sj}. This procedure of alternately estimating m_{sj} and a_{sk} was repeated until the residual sum of squares from the model estimating the a_{sk} parameters changed by less than 0.0001 between iterations. The results obtained using this method were compared to the results using the standard `nl` method to ensure they were identical. This process was not only quicker than the non-linear method, by a factor of about 30, but it was also easier to extend to robust methods.

This model was defined in a subset of our samples, the training set. This consisted of 70 subjects of each sex from each centre, chosen so that none of them had any vertebral deformities and none had any missing measurements. Excluding subjects with fractures is necessary since we wish to model normal vertebral heights (although an alternative to selecting subjects is presented in Chapter 6). However, it is possible that selecting subjects with no missing values could introduce bias in the measurements: for example, it may not be possible to visualize all of the spine of an unusually tall subject, and hence the tallest subjects may have been excluded from the training set. We need to test for this bias.

5.2.2. Fitting the Model to Subjects Not in Training Set

Once the model has been defined, the coefficients a_{sk} are known. Therefore predicting heights from a predefined model consists of fitting the regression model

$$H_{sij} = m_{sj} z_{si} + \epsilon_{sij}$$

where $z_{si} = \sum_{k=0}^{k=r} a_{sk} \times i^k$ is fixed. Note that there is no constant term in the regression model. This can be fitted in a very straightforward manner with standard statistical software. However, each subject needs their own magnification factor, so the number of parameters needed to be calculated can get very large. This is one reason why it was necessary to limit the size of the training set. However, once the a_{sk} are fixed, z_{si} is constant and takes the same value for each subject. Therefore, it is possible to fit the model to each subject individually, since the only parameter that still needs to be estimated is m_{sj}.

This makes it very simple to fit the model to the remaining subjects in each centre who were not included in the training set. Since there is likely to be some over-fitting in the training set, the fit of the model to the normal subjects in a given population who did not have any clinical fractures is a better indication of how well the model may fit other subjects taken from that population. This group of subjects without fractures but not used to define the model is referred to as the testing set.

This testing set (i.e. those normal subjects not used as part of the training set) can also be used to test for bias induced by excluding subjects with missing values from the training set. Subjects with missing values in the testing set could not have been included in the training set, whereas subjects without missing values could have been and were only excluded by chance. If we compare the residuals from subjects with missing values to those from subjects without missing values, we can test whether there was any bias introduced by excluding subjects with missing values from the training set.

5.2.3. Assessing the Fit of the Model

A good model of vertebral heights will both predict heights in normal subjects accurately and make it possible to discriminate between normal and fractured vertebrae.

The goodness of fit to normal vertebrae in the training set of polynomial models of different orders and with different numbers of magnification factors per subject was assessed using R^2, the proportion of the total variance of the vertebral heights that could be explained by the polynomial model. R^2 was also used to assess the goodness of fit of the chosen polynomial model to the testing set.

The same procedure could not be used to compare the polynomial models with the other models of interest, because it assumes that the mean of the residuals is zero. If the mean is not zero, R^2 will tend to exaggerate the goodness of fit of the model. As an extreme example, if the predicted height were always exactly 1mm greater than the measured height, the variance of the residuals would be 0, and R^2 would be 100%. Since the McCloskey-Kanis method tended to over-estimate the vertebral heights, R^2 would not be a good way of measuring goodness of fit for this model.

For this reason, the mean and the standard deviation of the residuals were used to compare different models. If the mean of the residuals is non-zero, it means that the method is biased, tending to consistently either under-estimate or over-estimate the true height. The standard deviation of the residuals gives a measure of the random error in the model.

The mean and standard deviation of the redsiduals from the chosen polynomial

model were compared to the mean and standard deviation of the residuals from the McCloskey-Kanis and Minne algorithms. This assessment was made for the training and testing samples separately. The distribution of the residuals compared using a Q-Q plot: if the residuals were normally distributed, the Q-Q plot should give a straight line.

It has been suggested that there are significant variations between populations in the shape of the spine. If this is the case, it may be that we may need to specify different models in different populations. On the other hand, it may be possible to formulate a model sufficiently general that the differences between populations can be accommodated by one or more parameters in the model, and the same model may be applied to all subjects. This would be a considerable advantage.

To get an idea of how serious the differences between populations are, models derived in one population were applied to subjects from a different population. The goodness of fit of the different models was then compared. Only subjects from the testing sample were used for this comparison.

5.2.4. Identification of Subjects with Deformities

There are a number of ways in which the above model can be used to identify subjects with deformities. One measure is the variance of the residuals: the within-subject variance should be similar in all subjects without deformities (due to measurement error and slight natural variation). However, if there is a fracture, the residual for that height will have a large negative value, and the variance of the residuals will consequently increase. Hence subjects with unusually large within-subject residual

variance may well have deformities.

Another method is to consider the largest negative residual. This has the advantage of identifying deformed *vertebrae*, not just *subjects* with deformed vertebrae. The largest negative residual may be measured in absolute terms (i.e. difference from expected height in mm) or in relative terms (i.e. % reduction from expected height). The latter method has the advantage of not being affected by magnification, whereas the former method is highly dependent on magnification.

In all three methods, a choice of threshold must be made such that all vertebrae (or subjects) on one side of the threshold are classed as normal and all those on the other side are classed as deformed. Two ways of defining this threshold were used. In the simpler method, the same threshold was used for each vertebra and site. The alternative involved calculating site-specific thresholds, which was done by calculating the mean and standard deviation of the residuals at each vertebra and site separately. The residual was converted to a z-score by subtracting the mean and dividing by the standard deviation. The threshold was then defined in terms of the z-score. The advantage of this method is that if a particular vertebra shows greater variation in shape, the standard deviation of the residuals will be greater, and therefore the height will have to be reduced to a greater extent in order for the vertebra to be classed as a fracture.

Two thresholds for each method were chosen for the purposes of comparison with other models: one which gave the same sensitivity as the reference model (McCloskey-Kanis) and one which gave the same specificity as the reference model in the training samples. However, if the model were to be accepted for routine use, it

is unlikely that a threshold chosen in this way would be used. Therefore, thresholds that may be of practical use were also tested.

As outlined in Chapter 3, the fractures we classified by the radiologist into one of the following types (in order of increasing severity): wedge, concavity, biconcavity, crush. The ability of each method to detect each of these types of deformity individually was therefore tested.

All three of the above methods were tested to detect individuals with deformities, and the latter two methods compared for identifying the location of the deformities. The methods were compared using receiver operating characteristic (ROC) curves. ROC curves are produced by varying the threshold at which a deformity is defined, and plotting the proportion of genuine fractures identified as fractures on the Y-axis against the proportion of normal vertebrae identified as fractures on the X-axis. The ability of each method to discriminate between fractures and normal vertebrae was measured by the area under the corresponding ROC curve. An area of 1 represents perfect discrimination, whilst an area of 0.5 represents no better discrimination than randomly allocating vertebrae as fractured or not.

In addition, logistic regression was used to compare the identification of fractured vertebrae. The logistic regression model calculates the probablity of a vertebra being fractured from the equation

$$\log\left(\frac{p}{1-p}\right) = a + bx \tag{5.2}$$

where p is the probability that the vertebra is fractured (or that the subject has at least one fractured vertebra), x is the smallest residual for that vertebra (or

subject) and a and b are coefficients. The pseudo-R^2 was calculated for each of these logistic regression models. There is no unique value of R^2 for logistic regression: Mittlbock and Schemper [19] have compared twelve different definitions, all with some, but not all, of the properties of R^2 in linear regression. The value used here was that produced automatically by stata, and recommended by Judge et al [20]. It is calculated as $1 - (L_1/L_0)$ where L_1 is the likelihood of the fitted model and L_0 is the likelihood of the constant-only model. It can range from 0 (for the constant-only model) to 1, if the predicted probability is 1 for all subjects with the outcome in question and 0 for all those without.

5.3. Results

5.3.1. Fit of Model To Training Sets

The fit of the various polynomial models to the training sets is summarised in tables 5.1 and 5.2. The fit of each model is assessed by R^2, the proportion of the variation in vertebral heights explained by the model. The table also gives an F-statistic and corresponding p-value for the improvement in the model due to adding an addition polynomial term, and the percentage reduction in the residual sum of squared achieved through adding the extra term.

It is clear that the cubic model is not a good fit: the residual mean squared error term can be reduced by between 7% - 18% by adding a quartic term. However, fitting any polynomial of order greater than 6 only gives a very slight, though possibly significant, improvement in fit. In addition, fitting the higher order models becomes

Population	Order	One Magnification Parameter				Three Magnification Parameters			
		R^2	%Change in RSS	F	p	R^2	%Change in RSS	F	p
Heidelberg	3	88.40%				89.2%			
	4	89.38%	8.4%	82.2	0.000	90.2%	9.1%	84.3	0.000
	5	89.63%	2.2%	20.9	0.000	90.4%	2.4%	21.5	0.000
	6	89.74%	0.9%	9.4	0.000	90.6%	1.0%	9.7	0.000
	7	89.79%	0.4%	4.5	0.003	90.6%	0.4%	4.7	0.003
	8	89.83%	0.2%	2.8	0.038	90.6%	0.2%	2.9	0.033
	9	89.84%	0.0%	1.0	0.389	90.6%	0.0%	1.0	0.38
	10	89.84%	0.0%			90.6%	0.0%		
Malmo	3	88.34%				89.0%			
	4	89.18%	7.1%	68.9	0.000	89.8%	7.6%	69.5	0.000
	5	89.48%	2.6%	24.9	0.000	90.1%	2.8%	25.2	0.000
	6	89.61%	1.1%	11.1	0.000	90.2%	1.2%	11.2	0.000
	7	89.65%	0.3%	3.5	0.016	90.3%	0.3%	3.5	0.015
	8	89.67%	0.1%	1.7	0.17	90.3%	0.1%	1.7	0.17
	9	89.68%	-0.1%	0.4	0.78	90.3%	-0.1%	0.4	0.78
	10	89.68%	0.0%			90.3%	0.0%		
Graz	3	91.73%					92.2%		
	4	92.91%	14.1%	145.9	0.000	93.4%	15.0%	148.9	0.000
	5	93.04%	1.8%	16.9	0.000	93.5%	1.9%	17.3	0.000
	6	93.30%	3.7%	34.5	0.000	93.8%	4.0%	35.4	0.000
	7	93.31%	-0.1%	0.3	0.86	93.8%	-0.1%	0.3	0.86
	8	93.32%	0.2%	2.4	0.06	93.8%	0.2%	2.5	0.06
	9	93.34%	0.1%	1.8	0.15	93.8%	0.1%	1.8	0.14
	10	93.34%	0.1%		0.	93.8%	0.1%		

Table 5.1.: Fit of polynomial models to male training sets

Population	Order	One Magnification Parameter				Three Magnification Parameters			
		R^2	%Change	F	p	R^2	%Change	F	p
Heidelberg	3	91.88%				92.5%			
	4	92.76%	10.8%	108.2	0.000	93.3%	11.6%	111.2	0.000
	5	92.85%	1.1%	10.5	0.000	93.4%	1.2%	10.9	0.000
	6	92.91%	0.8%	8.0	0.000	93.5%	0.9%	8.3	0.000
	7	92.95%	0.3%	4.1	0.007	93.5%	0.4%	4.2	0.006
	8	92.96%	0.0%	1.3	0.28	93.5%	0.0%	1.3	0.28
	9	92.96%	0.0%	0.8	0.47	93.5%	0.0%	0.9	0.45R
	10	92.97%	0.1%			93.5%	0.0%		
Malmo	3	90.51%				91.0%			
	4	91.37%	8.9%	87.5	0.000	91.9%	9.5%	88.5	0.000
	5	91.42%	0.5%	5.7	0.001	92.0%	0.6%	5.7	0.001
	6	91.47%	0.4%	4.9	0.002	92.0%	0.5%	5.0	0.002
	7	91.51%	0.3%	4.0	0.007	92.0%	0.4%	4.1	0.007
	8	91.54%	0.2%	3.1	0.025	92.1%	0.3%	3.1	0.025
	9	91.54%	-0.1%	0.3	0.84	92.1%	-0.1%	0.3	0.84
	10	91.54%	0.0%		0.631	92.1%	0.0%		
Graz	3	91.49%				92.0%			
	4	92.96%	17.2%	184.0	0.000	93.5%	18.3%	188.1	0.000
	5	92.99%	0.3%	3.9	0.009	93.5%	0.4%	4.0	0.007
	6	93.16%	2.2%	21.2	0.000	93.7%	2.4%	21.8	0.000
	7	93.17%	0.1%	1.8	0.15	93.7%	0.1%	1.8	0.14
	8	93.18%	0.1%	1.7	0.18	93.7%	0.1%	1.7	0.16
	9	93.19%	-0.1%	0.4	0.76	93.7%	-0.1%	0.4	0.75
	10	93.19%	0.0%			93.7%	0.0%		

Table 5.2.: Fit of polynomial models to female training sets

difficult due to numerical instabilities: the a_{sk} coefficients are extremely small and the i^k are extremely large.

The models of orders 9 and 10 could not be fitted using the iterative method, and had to be fitted using conventional non-linear methods. In addition, the models of order 10 were not identified: the degrees of freedom reported by the `nl` procedure in stata were the same for models of orders 9 and 10, and standard errors were not calculated for three of the parameters of the 10^{th} order model. For this reason, F statistics cannot be calculated for these models.

The improvement in fit achieved by using three separate magnification factors rather than 1 is illustrated in tables 5.3 and 5.4. Only the F, p and % improvement figures are given, since the R^2 values are in Tables 5.1 and 5.2.

The degree to which fitting 3 magnification factors rather than 1 improves the fit of the model varies between the training sets. In men and women from Malmo, it is of the order of 1%, and of borderline statistical significance at best. However, in the men and women from Graz and Heidelberg, it was around 2% - 3% and highly statistically significant.

Clearly, the 'normal' shape of the spine varies between these populations. However, a 6th order polynomial with three magnification factors seems to provide a good fit in all the populations, so this model was explored further.

Comparison with Other Models

For the sake of comparison, the residual sum of squares (RSS) of the Minne and McCloskey-Kanis models in these subjects is summarised in table 5.5, along with the

Population	Order	%Change	F	p
Heidelberg	3	1.9	1.4	0.004
	4	2.5	1.5	0.000
	5	2.7	1.5	0.000
	6	2.8	1.6	0.000
	7	2.8	1.6	0.000
	8	2.9	1.6	0.000
	9	2.9	1.6	0.000
	10	2.9	1.6	0.000
Malmo	3	0.2	1.0	0.36
	4	0.7	1.1	0.15
	5	0.8	1.2	0.10
	6	0.9	1.2	0.08
	7	0.9	1.2	0.08
	8	0.9	1.2	0.08
	9	0.9	1.2	0.08
	10	0.9	1.2	0.08
Graz	3	0.9	1.2	0.08
	4	2.0	1.4	0.001
	5	2.1	1.4	0.001
	6	2.4	1.5	0.000
	7	2.4	1.5	0.000
	8	2.5	1.5	0.000
	9	2.5	1.5	0.000
	10	2.5	1.5	0.000

Table 5.3.: Improvement due to fitting three magnification factors rather than one in male training sets

Population	Order	%Change	F	p
Heidelberg	3	2.1	1.4	0.001
	4	3.0	1.6	0.000
	5	3.1	1.6	0.000
	6	3.2	1.6	0.000
	7	3.2	1.6	0.000
	8	3.2	1.6	0.000
	9	3.2	1.6	0.000
	10	3.2	1.6	0.000
Malmo	3	0.4	1.1	0.26
	4	1.0	1.2	0.07
	5	1.0	1.2	0.06
	6	1.1	1.2	0.05
	7	1.1	1.2	0.05
	8	1.1	1.2	0.05
	9	1.1	1.2	0.05
	10	1.1	1.2	0.05
Graz	3	0.9	1.2	0.09
	4	2.2	1.4	0.001
	5	2.2	1.4	0.001
	6	2.4	1.5	0.000
	7	2.4	1.5	0.000
	8	2.4	1.5	0.000
	9	2.4	1.5	0.000
	10	2.4	1.5	0.000

Table 5.4.: Improvement due to fitting three magnification factors rather than one in female training sets

RSS from the 3^{rd} and 6^{th} order polynomial models with 3 magnification factors. It can be seen that the residual sum of squares is approximately 2 to 3 times as great with the Minne model as it is with the polynomial model with 3 degrees of freedom. Given that the degrees of freedom of the Minne is the same as the cubic model, this shows a much poorer fit.

The situation is less clear-cut with the Mc-K model. In one case the RSS is greater than the cubic model, in others it is less but greater than the 6th order model. However, it uses considerably fewer degrees of freedom than either of these models, so the RSS cannot be used directly as a criterion for comparison. However, if we assume that the Mc-K model requires 39 parameters (for the H_a/H_p, H_m/H_p and H_p/H_{pp} ratios need to be estimated for each vertebral level), we can still say that the 6th order polynomial model fits significantly better to these subjects.

Set	Polynomial Models			
	3rd Order	6th Order	Minne	McCloskey-Kanis
Heidelberg Men	6371.5	5577.5	16287.5	7743.7
Heidelberg Women	4449.7	3838.6	10534.8	4519.1
Malmo Men	6510.3	5757.3	15155.4	7002.3
Malmo Women	5287.3	4720.6	11650.2	5974.1
Graz Men	4580.9	3654.5	9160.3	4234.2
Graz Women	4718.1	3735.7	9564.5	4430.8

Table 5.5.: Residual Sums of Squares in Training Sets

The distribution of the residuals, shown in Figure 5.1 illustrates further diffe-

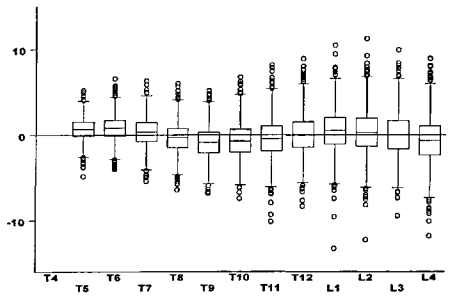
rences between the models.

It can be seen from figure 5.1(a) that the mean residual differs between the different vertebral levels. This shows that a cubic polynomial is not sufficient to model the natural shape of the spine.

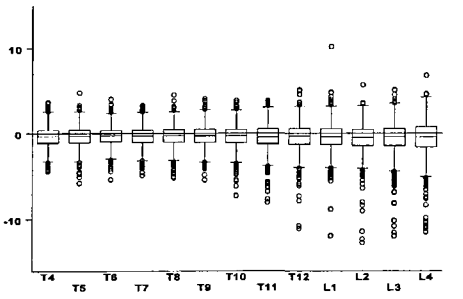
Although there is less variation between vertebral levels in the mean residual with the McCloskey-Kanis model, the mean residual is less than zero for all vertebrae. In other words, this estimator is biased, consistently predicting vertebral heights to be greater than they should be. The reason why we might have expected this to be the case is given in section 2.3.5. It can also be seen that there are a number of heights predicted to be very much larger than their measured heights, which are not apparent using either of the other models.

There is slight variation between vertebrae in the mean residual with the polynomial model, but not nearly as great as with the Minne model. The mean residual is nearer to zero for this estimator (i.e. it is less biased) and the variance of the residuals is smaller (i.e. it is more precise). Thus, at least in the training sets, the polynomial model provides a better fit than either of the other models.

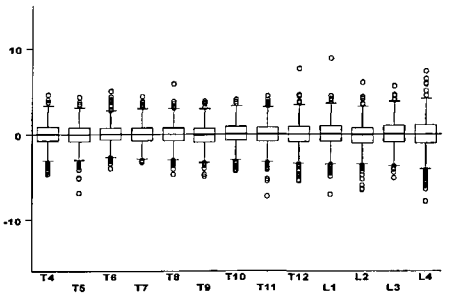
Figure 5.2 shows q-q plots of the residuals from the three methods. It can be seen that in all methods, the residuals are not normally distributed, with an excess of both very small and very large residuals. However, the excess of very small residuals is greater in the McCloskey-Kanis method than the other two.



(a) Minne Model

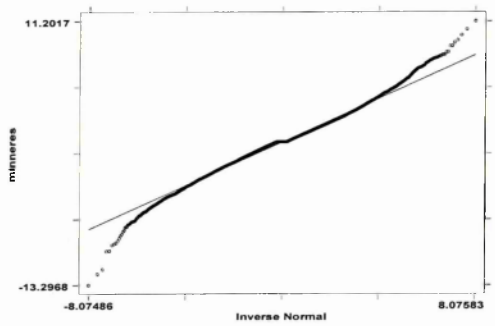


(b) McCloskey-Kanis Model

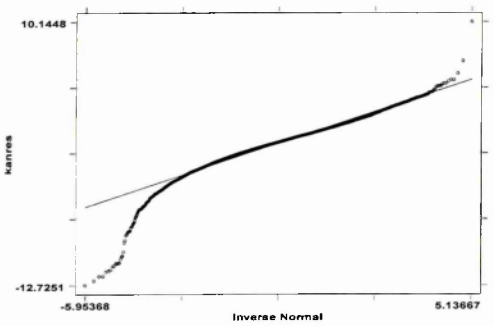


(c) 6^{th} Order Polynomial Model

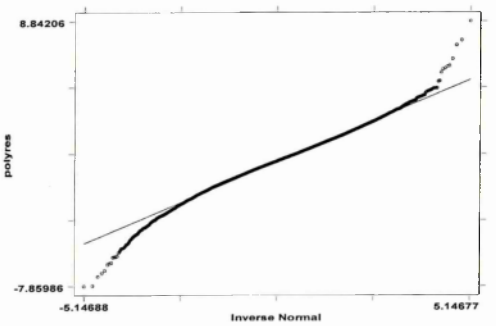
Figure 5.1.: Distribution of residuals from morphometric models



(a) Minne Model



(b) McCloskey-Kanis Model



(c) 6^{th} Order Polynomial Model

Figure 5.2.: Normal plots of residuals from morphometric models

5.3.2. Fit of Models To Other Normal Subjects

The fit of the models to the data used in their definition is not the ideal measure of the adequacy of the models. The models may 'overfit' this data, and the fit of the models to a different sample from the same population can be expected to be less good. It may be that the bias in the estimation of the goodness of fit differs between the different methods, the fit to the training set is therefore misleading. The models were therefore also fitted to the 'testing' sets: the subjects without clinically apparent deformities who were not used in the model definition stage.

For comparing the fit of the models in other subjects, of the polynomial models, only the 6^{th} order model with 3 magnification factors was considered. Table 5.6 shows the RSS for this polynomial model, as well as the Minne and McCloskey-Kanis models to the 'testing' samples in each centre.

Population	RSS		
	Polynomial	Minne	McCloskey-Kanis
Heidelberg Men	4108.9	11005.0	4461.3
Heidelberg Women	3261.4	7786.0	4450.2
Malmo Men	6898.4	18670.6	8626.9
Malmo Women	6967.5	15618.8	9467.5
Graz Men	4949.7	11534.2	5745.8
Graz Women	5681.8	14822.5	6366.7

Table 5.6.: Residual sums of squares of morphometric models in testing sample

The fit of the Minne model is less good then either of the other two models. The fit of the polynomial model looks equal to, or slightly better than, the fit of the McCloskey-Kanis model. However, the McCloskey-Kanis model is again biased, with the predicted heights tending to be be bigger than the measured heights. Since R^2 does not take bias into account, it will overestimate the goodness of fit of the McCloskey-Kanis model. A better measure of the fit is given by the mean and standard deviation of the residuals: the mean giving an estimate of the bias and the standard deviation an estimate of the imprecision. These are given in table 5.7.

Population	Polynomial Model		Minne Model		McCloskey-Kanis Model	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Heidelberg Men	-0.01	1.40	0.39	2.34	-0.45	1.46
Heidelberg Women	0.02	1.16	-0.29	1.88	-0.63	1.35
Malmo Men	0.02	1.50	-0.42	2.54	-0.60	1.68
Malmo Women	-0.03	1.38	0.49	2.14	-0.56	1.61
Graz Men	0.00	1.27	0.20	2.06	-0.45	1.37
Graz Women	0.02	1.17	-0.15	1.96	-0.32	1.24

Table 5.7.: Mean and standard deviation of residuals in testing subgroups

It is clear from table 5.7 that the bias is least for the polynomial method. There are considerable biases with the Minne method, although sometimes the heights are systematically overestimated and sometimes they are systematically underestimated. Again, the McCloskey-Kanis method overestimates heights by between 0.3mm and

0.6mm.

The standard deviation of the residuals is again greatest with the Minne method. There is very little difference between the other methods, but the standard deviation using the polynomial method is consistently lower.

5.3.3. Effect of Using a Different Population

It has been shown that for the Minne and McCloskey-Kanis methods, each population needs to have its own reference range defined. It is not a priori obvious that the same is true for the polynomial models. We looked at two different ways of applying a polynomial model defined in a different population:

1. Using subjects of the same sex from different centres.

2. Using subjects of the opposite sex from the same centre.

In both cases, a sixth order polynomial model was used. The results of these comparisons are shown in tables 5.8 and 5.9.

It can be seen that using the opposite gender for defining the model is unsuccessful. The imprecision of the estimated heights is greater, and the magnitude of the bias is also greater. There is a tendency to underestimate the heights in men if a female population is used to define the model, and to overestimate the heights in women using a male reference range. However, the bias is less than 0.2mm in all cases, so it is not too extreme.

However, using a model derived in a different centre appears to make little difference to the accuracy of the predictions. The bias is consistently small, and the

Population	Same Sex		Opposite Sex	
	Mean	S.D.	Mean	S.D.
Heidelberg Men	0.01	1.45	-0.10	1.66
Heidelberg Women	-0.01	1.23	0.07	1.36
Malmo Men	-0.01	1.58	-0.14	1.87
Malmo Women	0.03	1.46	0.08	1.61
Graz Men	0.00	1.32	-0.10	1.51
Graz Women	-0.02	1.21	0.06	1.32

Table 5.8.: Mean and standard deviation of residuals in testing subgroups using reference of opposite gender

Population	Heidelberg		Malmo		Graz	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Heidelberg Men	0.01	1.45	0.02	1.46	-0.03	1.51
Heidelberg Women	-0.01	1.23	0.02	1.25	-0.05	1.29
Malmo Men	-0.02	1.57	-0.01	1.58	-0.06	1.65
Malmo Women	-0.01	1.45	0.03	1.46	-0.05	1.52
Graz Men	0.03	1.41	0.04	1.40	0.00	1.32
Graz Women	0.01	1.27	0.04	1.26	-0.02	1.21

Table 5.9.: Mean and standard deviation of residuals in testing subgroups using all references

Sample	Subjects with Missing Data
Heidelberg Men	13/126 = 10%
Heidelberg Women	10/134 = 7%
Malmo Men	15/150 = 10%
Malmo Women	17/166 = 10%
Graz Men	27/153 = 18%
Graz Women	17/180 = 9%

Table 5.10.: Numbers of deformity-free subjects with and without missing data

imprecision only increases slightly if a different centre is used to develop the model. However, there is a suggestion that Graz behaves slightly differently from the other centres: the imprecision increases if Graz is used as a model for the other centres, or if the other centres are used as a model for Graz. Nonetheless, the standard deviation of the residuals is in general smaller than that using the McCloskey-Kanis model, even if a polynomial defined in a different population is used.

5.3.4. Effect of Excluding Subjects with Missing Data From the Training Set

The numbers of subjects of each sex in each centre with at least one missing height measurement are given in Table 5.10. In most of the samples, about 10% of subjects had at least one missing measurement.

There were a total of 189 vertebrae that could not be measured. A disproportionate number of these were at either T4 or T5, is shown in Table 5.11.

vertebra	Missing Measurements	
	Number	Percentage
T4	44	23.3
T5	25	13.2
T6	12	6.4
T7	9	4.8
T8	10	5.3
T9	13	6.9
T10	18	9.5
T11	12	6.4
T12	15	7.9
L1	2	1.1
L2	4	2.1
L3	11	5.8
L4	14	7.4

Table 5.11.: Numbers of vertebrae with missing data at each vertebral level

The mean residual does not differ significantly between those with and without missing values (0.01mm vs 0.00mm), but the standard deviation of the residuals is slightly greater in the subjects with missing data (1.40mm vs 1.29mm $p < 0.001$ using an F-test). This difference is not due to outlying observations, since the most extreme observations are in subjects without missing data. However, Table 5.10 shows that men are more likely to have missing data, and also have greater variation in vertebral height. If we stratify by gender, the difference in standard deviation is no longer significant in the women (1.26mm vs 1.23mm, $p = 0.27$), but remains significant in the men (1.49mm vs 1.36mm, $p < 0.001$).

5.3.5. Identification of Deformed Vertebrae

Treatment of Residuals

Table 5.12 shows the pseudo-R^2 values and area under the ROC curve for logistic regression models in which the outcome is the presence of a fracture, and each model contains a single predictor variable, being one of the the four possible treatments of residuals:

1. The largest absolute reduction reduction in height in any of the three heights in that vertebra;

2. The largest percentage reduction in height in any of the three heights in that vertebra;

3. The ratio of the largest reduction in height in any of the three heights in that vertebra to the standard deviation of the residuals for that vertebra and site

(to give a site specific threshold for absolute height reduction);

4. The ratio of the largest percentage reduction in height in any of the three heights in that vertebra to the standard deviation of the percentage differences between the measured and predicted heights for that vertebra and site (to give a site specific threshold for percentage height reduction).

These values were calculated for all three models.

It is clear from the above table that relative residuals perform better than absolute residuals. Whether a site specific threshold should be used is more difficult to say, since such thresholds work better with the absolute residuals, but only in the Minne model with relative residuals. Note that the Mcloskey-Kanis model appears to discriminate better between fractured and unfractured vertebrae, despite not fitting as well to the unfractured subjects. Also, the polynomial model with a single magnification factor per subject discriminates better between fractured and unfractured vertebrae, despite not fitting as well to the unfractured heights. The reason why this might be the case is explored in Section 5.4.

McCloskey-Kanis Method

The numbers of vertebrae identified as deformities in each subgroup of each population using the McCloskey-Kanis method are given in table 5.13 below. Within each subject who had at least one fracture, there were also unfractured vertebrae, so the fractured vertebrae and normal vertebrae were analysed separately, since we need to be able to discriminate between fractured and unfractured vertebrae even

	R^2				AUC			
	McCloskey-Kanis	Minne	Poly		McCloskey-Kanis	Minne	Poly	
			3 mag. factors	1 mag. factor			3 mag. factors	1 mag. factor
Absolute Residuals, Common Threshold	0.5555	0.3907	0.4919	0.5489	0.9872	0.9080	0.9539	0.9771
Relative Residuals, Common Threshold	0.6626	0.4800	0.5481	0.6096	0.9945	0.9301	0.9629	0.9856
Absolute Residuals, Site Specific Threshold	0.6025	0.4648	0.5038	0.5723	0.9898	0.9380	0.9535	0.9784
Relative Residuals, Site Specific Threshold	0.6198	0.4925	0.5167	0.5794	0.9906	0.9443	0.9572	0.9806

Table 5.12.: Goodness of prediction of different functions of residuals

in subjects with fractures.

Less than 1% of vertebrae in the training set and less than 2% of vertebrae in the other normal subjects were classed as deformities using this method. Interestingly, in subjects with fractures, the proportion of normal vertebrae classed as deformities was higher. We would expect rather more false positives in the subjects known to have deformities with causes other than fracture, since a morphometric method can not determine the cause of the deformity, but in fact the rates are not excessively high. There is considerable variation between centres in the proportion of fractures classed as deformities using this method.

Minne Method

The numbers of vertebrae identified as deformities in each subgroup of each population using the Minne method are given in table 5.14 below.

This model performs less well than the McCloskey-Kanis model. The false positive rate is considerably higher in all groups of normal vertebrae. In addition, fewer vertebrae can be evaluated using this method, since if the measurements for T4 are missing, no vertebrae in that spine can be evaluated. The sensitivity is slightly greater for some centre-sex combinations, but the absolute number of fractures detected is smaller since fewer vertebrae can be evaluated.

Polynomial Method: Same Specificity as McCloskey-Kanis

From Table 5.12, it appears that the best way to define deformities from the polynomial model is to use the relative reduction in height. Since there were 34 false

Population	Training Set	Other Normals	Fractures		Other Deformities
	(All Verts)	(All Verts)	(Norm. Verts)	(Fx Verts)	(All Verts)
Heidelberg Men	7/910=0.8%	0/701=0.0%	3/271=1.1%	18/35=51.4%	7/476=1.5%
Heidelberg Women	0/910=0.0%	6/811=0.7%	7/208=3.4%	48/49=98.0%	2/140=1.4%
Malmo Men	6/910=0.7%	14/1022=1.4%	16/439=3.6%	46/61=75.4%	22/209=10.5%
Malmo Women	9/910=1.0%	17/1218=1.4%	17/450=3.8%	62/62=100.0%	11/114=9.7%
Graz Men	3/910=0.3%	12/1025=1.2%	10/389=2.6%	45/54=83.3%	9/242=3.3%
Graz Women	9/910=1.0%	8/1391=0.6%	3/225=1.3%	21/30=70.0%	3/104=2.9%
Total	34/5460 = 0.6%	57/6168 = 0.9%	56/1982=2.8%	240/291=82.5%	54/1285=4.2%

Table 5.13.: Number of McCloskey-Kanis deformities

Population	Training Set	Other Normals	Fractures		Other Deformities
	(All Verts)	(All Verts)	(Norm. Verts)	(Fx Verts)	(All Verts)
Heidelberg Men	29/910=3.2%	20/671=3.0%	13/264=4.9%	18/32=56.3%	15/364=4.1%
Heidelberg Women	32/910=3.5%	26/736=3.5%	15/203=7.4%	39/43=90.7%	5/123=4.1%
Malmo Men	22/910=2.4%	47/964=4.9%	38/397=9.6%	44/55=80.0%	32/203=15.8%
Malmo Women	25/910=2.7%	43/1132=3.8%	33/440=7.5%	49/61=80.3%	12/102=11.8%
Graz Men	30/910=3.3%	32/903=3.5%	24/367=6.5%	38/48=79.2%	11/218=5.1%
Graz Women	19/910=2.1%	45/1286=3.5%	7/206=3.4%	18/27=66.7%	6/92=6.5%
Total	157/5460=2.9%	213/5692=3.7%	130/1877=6.9%	206/266=77.4%	81/1102=7.4%

Table 5.14.: Number of Minne deformities

positive deformities in the training sets using the McCloskey-Kanis method, we chose the threshold to give the same number of false positives using this method (i.e. half-way between the 34^{th} smallest and 35^{th} smallest values in the training set).

Using three magnification factors per subject and a common threshold for all sites, the threshold chosen in this way was -15.7%: i.e. if any vertebral height was more than 15.7% less than its predicted value, that vertebra would be classed as a deformity. For the site-specific threshold method, the threshold was at a z-score of -3.20, corresponding to a reduction of between 12.2% and 23.2%. Using a single magnfication factor per subject, the corresponding thresholds were 15.7% and -3.24, which gave a reduction of between 13.2% and 22.4%.The actual numbers of vertebrae classed as deformities in this way is given in Table 5.15 for the use of three magnification factors and Table 5.16 for the use of a single magnification factor.

There are similar numbers of false positives in the testing sets using this method compared to the McCloskey-Kanis method, and fewer false positives in the unfractured vertebrae in subjects with fractures. However, the sensitivity of this method is slightly less than that of the McCloskey-Kanis method. The best option appears to be to use a single magnification factor and a common threshold: this detects 229 fractures, compared to 240 for the McCloskey-Kanis model.

Polynomial Method: Same Sensitivity as McCloskey-Kanis

The McCloskey-Kanis model correctly identified 240 of the 291 vertebral fractures identified by the clinician. Therefore, to obtain the same sensitivity, the threshold chosen for the polynomial model should also identify 240 true fractures. Using a

Population	Site-Specific	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	No	14/910=1.3%	4/701=0.6%	4/271=1.5%	18/35=51.4%	8/476 =1.7%
Heidelberg Men	Yes	11/910=1.2%	5/701=0.7%	4/271=1.5%	13/35=37.1%	5/476=1.1%
Heidelberg Women	No	1/910=0.1%	2/811=0.3%	4/208=1.9%	38/49=77.6%	3/139=2.2%
Heidelberg Women	Yes	4/910=0.4%	7/811=0.9%	5/208=2.4%	37/49=75.5%	6/139=4.3%
Malmo Men	No	5/910=0.6%	8/1022=0.8%	6/439=1.4%	46/61=75.4%	18/209=8.6%
Malmo Men	Yes	5/910=0.6%	10/1022=1.0%	10/439=2.3%	44/61=72.1%	19/209=9.1%
Malmo Women	No	6/910=0.7%	12/1218=1.0%	7/450=1.6%	54/62=85.7%	6/114=5.3%
Malmo Women	Yes	7/910=0.8%	18/1218=1.5%	7/450=1.6%	51/62=82.3%	8/114=7.0%
Graz Men	No	3/910=0.3%	8/1025=0.8%	2/389=0.5%	36/54=66.7%	4/241=1.7%
Graz Men	Yes	5/910=0.5%	15/1025=1.5%	4/389=1.0%	37/54=68.5%	8/241=3.3%
Graz Women	No	5/910=0.6%	4/1391=0.3%	1/225=0.4%	19/30=63.3%	1/104=1.0%
Graz Women	Yes	4/910=0.4%	9/1391=0.7%	3/225=1.3%	19/30=63.3%	2/104=1.9%
Total	No	34/5460 = 0.6%	38/6168 = 0.6%	24/1982=1.2%	211/291=71.8%	40/1283=3.1%
Total	Yes	34/5460 = 0.6%	64/6168 = 1.0%	33/1982=1.7%	201/291=69.1%	48/1283=3.7%

Table 5.15.: Number of deformities from polynomial model with three magnification factors and the same specificity as McCloskey-Kanis model in training samples

Population	Site-Specific	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	No	12/910=1.3%	4/701=0.6%	6/271=2.2%	20/35=57.1%	12/476 =2.5%
Heidelberg Men	Yes	10/910=1.1%	4/701=0.6%	4/271=1.5%	16/35=45.7%	6/476=1.3%
Heidelberg Women	No	2/910=0.2%	2/811=0.3%	5/208=2.4%	43/49=87.8.6%	3/139=2.2%
Heidelberg Women	Yes	3/910=0.3%	5/811=0.6%	6/208=2.9%	42/49=85.7%	6/139=4.3%
Malmo Men	No	5/910=0.6%	11/1022=1.1%	8/439=1.8%	48/61=78.7%	19/209=9.1%
Malmo Men	Yes	5/910=0.6%	15/1022=1.5%	7/439=1.6%	44/61=72.1%	17/209=8.1%
Malmo Women	No	6/910=0.7%	12/1218=1.0%	7/450=1.6%	54/62=87.1%	6/114=5.3%
Malmo Women	Yes	6/910=0.7%	17/1218=1.4%	7/450=1.6%	55/62=88.7%	10/114=8.8%
Graz Men	No	4/910=0.4%	12/1025=1.2%	2/389=0.5%	44/54=81.5%	6/241=2.5%
Graz Men	Yes	5/910=0.6%	12/1025=1.2%	4/389=1.0%	42/54=77.8%	7/241=2.9%
Graz Women	No	5/910=0.6%	7/1391=0.5%	1/225=0.4%	20/30=66.7%	1/104=1.0%
Graz Women	Yes	5/910=0.6%	9/1391=0.7%	1/225=0.4%	20/30=66.7%	1/104=1.0%
Total	No	34/5460 = 0.6%	49/6168 = 0.8%	29/1982=1.5%	229/291=78.7%	49/1283=3.8%
Total	Yes	34/5460 = 0.6%	62/6168 = 1.0%	29/1982=1.5%	219/291=75.3%	47/1283=3.7%

Table 5.16.: Number of deformities from polynomial model with one magnification factor and the same specificity as McCloskey-Kanis model in training samples

common threshold for all sites and three magnification factors, the threshold chosen in this way was -12.3%. For the site-specific threshold method, the threshold was at a z-score of -2.55, corresponding to a reduction of between 9.7% and 18.5%. Using a single magnfication factor per subject, the corresponding thresholds were 14.9% and -2.82, which gave a reduction of between 11.5% and 19.5%. The actual numbers of vertebrae classed as deformities in this way are given in Tables 5.17 and 5.18.

Identification of Deformity Types

Table 5.19 gives the number of fractures of each of the 4 types correctly identified by each of the methods being compared. Since sensitivities are being compared, the polynomial model with the same specificity as the McCloskey-Kanis model was used. The polynomial model with a single magnification factor was used, since it performed better than having three magnification factors.

Clearly, the more severe deformities are more likely to be detected than the less severe ones. Since wedge deformities make up a greater proportion of the fractures in men than in women, we might expect the morphometric methods to be less sensitive overall to fractures in men than in women. Table 5.20 compares the overall performance of each of the methods in men and women, and shows that morphometry is less sensitive in men, whichever method is used.

5.3.6. Identification of Subjects with Deformities

Table 5.21 shows the classification of subjects as cases or non-cases using the McCloskey-Kanis and Minne methods, according to whether or not they had vertebral fractures

Population	Site-Specific	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	No	38/910=4.2%	12/701=1.7%	12/271=4.4%	23/35=65.7%	23/476 =4.8%
Heidelberg Men	Yes	29/910=3.2%	11/701=1.6%	9/271=3.3%	21/35=60.0%	15/476=3.2%
Heidelberg Women	No	11/910=1.2%	10/811=1.2%	9/208=4.3%	39/49=79.6%	6/139=4.3%
Heidelberg Women	Yes	20/910=2.2%	16/811=2.0%	12/208=5.8%	39/49=79.6%	13/139=9.4%
Malmo Men	No	18/910=2.0%	25/1022=2.5%	15/439=3.4%	52/61=82.3%	23/209=11.0%
Malmo Men	Yes	13/910=1.4%	39/1022=3.8%	20/439=4.5%	51/61=83.6%	25/209=12.0%
Malmo Women	No	14/910=1.5%	46/1218=3.8%	14/450=3.1%	56/62=90.3%	9/114=7.9%
Malmo Women	Yes	18/910=2.0%	49/1218=4.0%	18/450=4.0%	56/62=90.3%	14/114=12.3%
Graz Men	No	11/910=1.2%	24/1025=2.3%	7/389=1.8%	44/54=81.5%	9/241=3.7%
Graz Men	Yes	22/910=2.4%	36/1025=3.5%	13/389=3.3%	46/54=85.2%	11/241=4.6%
Graz Women	No	16/910=1.8%	29/1391=2.1%	3/225=1.3%	26/30=86.7%	3/104=2.9%
Graz Women	Yes	22/910=2.4%	41/1391=3.0%	5/225=2.2%	27/30=90.0%	6/104=5.8%
Total	No	108/5460 = 2.0%	146/6168 = 2.4%	60/1982=3.0%	240/291=82.5%	73/1283=5.7%
Total	Yes	124/5460 = 2.3%	192/6168 = 3.1%	77/1982=3.9%	240/291=82.5%	84/1283=6.6%

Table 5.17.: Number of deformities from polynomial model with three magnification factors and the same sensitivity as McCloskey-Kanis model in training samples

Population	Site-Specific	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	No	18/910=2.0%	6/701=0.9%	7/271=2.6%	21/35=60.0%	13/476 =2.7%
Heidelberg Men	Yes	19/910=2.1%	7/701=1.0%	4/271=1.5%	20/35=57.1%	12/476=2.5%
Heidelberg Women	No	4/910=0.4%	3/811=0.4%	6/208=2.9%	43/49=87.8%	4/139=2.9%
Heidelberg Women	Yes	11/910=1.2%	11/811=1.4%	8/208=3.9%	43/49=87.8%	10/139=7.2%
Malmo Men	No	6/910=0.7%	15/1022=1.5%	10/439=2.3%	50/61=82.0%	20/209=9.6%
Malmo Men	Yes	8/910=0.9%	27/1022=2.6%	13/439=3.0%	54/61=88.5%	21/209=10.1%
Malmo Women	No	8/910=0.9%	15/1218=1.2%	9/450=2.0%	56/62=90.3%	8/114=7.0%
Malmo Women	Yes	11/910=1.2%	32/1218=2.6%	14/450=3.1%	56/62=90.3%	12/114=10.5%
Graz Men	No	4/910=0.4%	14/1025=1.4%	3/389=0.8%	47/54=87.0%	6/241=2.5%
Graz Men	Yes	11/910=1.2%	28/1025=2.7%	11/389=2.8%	45/52=83.3%	10/241=4.2%
Graz Women	No	9/910=1.0%	11/1391=0.8%	1/225=0.4%	23/30=76.7%	1/104=1.0%
Graz Women	Yes	12/910=1.3%	23/1391=1.7%	3/225=1.3%	22/30=73.3%	5/104=4.8%
Total	No	49/5460 = 0.9%	64/6168 = 1.0%	36/1982=1.8%	240/291=82.5%	52/1283=4.1%
Total	Yes	72/5460 = 1.3%	129/6168 = 2.1%	53/1982=2.7%	240/291=82.5%	70/1283=5.5%

Table 5.18.: Number of deformities from polynomial model with one magnification factor and the same sensitivity as McCloskey-Kanis model in training samples

Fracture Type	Number of Fractures	Polynomial Model		McCloskey-Kanis	Minne
		Common Threshold	Site-Specific Threshold		
Wedge	160	67.5%	61.9%	69.4%	66.7%
Concave	101	90.1%	91.1%	99.0%	42.6%
Biconcave	18	100%	100%	100%	64.3%
Crush	7	100%	100%	100%	100%

Table 5.19.: Sensitivities of different methods to different types of fracture

Method	Women		Men	
	Sensitivity	Specificity	Sensitivity	Specificity
McCloskey-Kanis	93%	98.8%	73%	98.6%
Minne	81%	96.2%	74%	95.6%
Poly 15% Threshold	84.0%	99.0%	75.5%	98.5%
Poly 20% Threshold	70.1%	99.8%	48.4%	99.6%

Table 5.20.: Comparison of different morphometric methods in men and women

on the clinical reading of the films. The table also shows the classification of subjects as cases or non-cases using the polymomial model, using several different thresholds: 15%, 20%, 15.5% (to give the same sensitivity as the McCloskey-Kanis method) and 16.3% (to give the same specificity as the McCloskey-Kanis method).

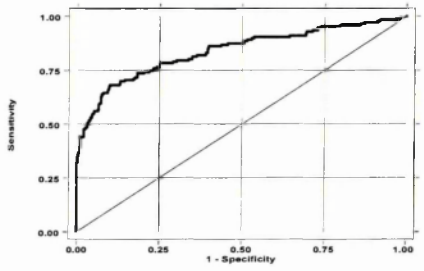
Using a threshold of 15% gives a method that is more sensitive than the McCloskey-Kanis method, but less specific, whilst using a 20% threshold is less sensitive but more specific. Choosing a threshold to give the same sensitivity or specificity in the training set as the McCloskey-Kanis method leads to very similar numbers of fractures. The Minne method is equally sensitive, but far less specific with nearly 1/4 of fracture-free subjects being false positives.

The ROC curves for five methods of identifying subjects with vertebral deformities are given in figure 5.3. The estimated areas and their standard errors are given in table 5.22. However, note that the areas cannot be compared using the standard errors as shown, since the methods were applied to the same subjects and hence are not independent. Formal testing of the areas, using the methods described by DeLong, DeLong and Clarke-Pearson [21] for comparing the areas of correlated ROC curves shows that the area is slightly, but not significantly greater for the McCloskey-Kanis method than for the polynomial model using relative residuals. Both of these methods are significantly better than using the standard deviation of the residuals, which is in turn significantly better than using absolute residuals. The AUC for the Minne model was significantly less than for any other model.

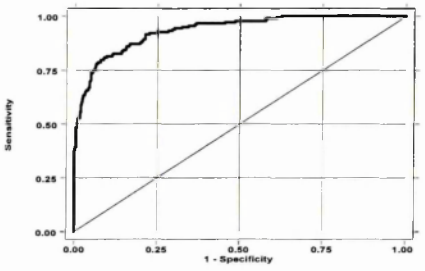
The differences in the performances of the different models in men and women were measured using logistic regression. ROC curves were calculated for men and

Method		Training Set	Testing Set	Fractures	Other Deformities
McCloskey-Kanis		23/420 = 5.5%	43/446 = 8.8%	152 / 178 = 85.4%	35/105 = 33.3%
Minne		102/420 = 24.3%	114/445 = 25.6%	140 / 166 = 84.3%	37/87 = 42.5%
Polynomial	15% Threshold	44/420=10.5%	51/489 = 10.4%	156 / 178 = 87.6%	39/105 = 37.1%
Polynomial	20% Threshold	9/420=2.1%	4/489 = 0.8%	122 / 178 = 68.5%	21/105 = 20.0%
Polynomial	15.5% Threshold	36/420=8.6%	48/489 = 9.8%	152 / 178 = 85.4%	39/105 = 37.1%
Polynomial	16.3% Threshold	23/420=5.5%	32/489 = 6.5%	147 / 178 = 82.6%	36/105 = 34.3%

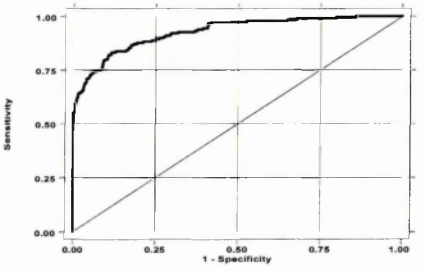
Table 5.21.: Cross-tabulation of morphometric and clinical classifications of the presence of at least one deformity in a subject



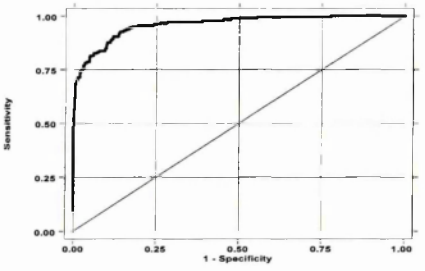
(a) Minne Model



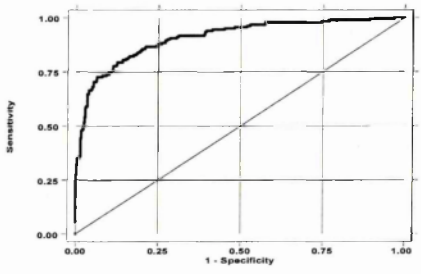
(b) McCloskey-Kanis Model



(c) Polynomial Model (Absolute Residuals)



(d) Polynomial Model (Relative Residuals)



(e) Polynomial Model (Variance of Residuals)

Figure 5.3.: ROC curves for various morphometric models

Model	Area	Standard Error
McCloskey-Kanis Model	0.9597	0.0089
Polynomial Model (Relative Residuals)	0.9480	0.0067
Polynomial Model (Variance of Residuals)	0.9324	0.0143
Polynomial Model (Absolute Residuals)	0.9193	0.0120
Minne Model	0.7915	0.0197

Table 5.22.: Areas under ROC curves

women separately, and the interaction between gender and and the model parameter used to determine if the effect differed significantly. The results are in table 5.23.

Method	AUC in Men	AUC in Women for difference	p-value
McCloskey-Kanis	0.9390	0.9785	0.004
Minne	0.8112	0.8585	0.031
Polynomial (Relative Residuals)	0.9317	0.9638	0.25
Polynomial (Absolute Residuals)	0.8986	0.9330	0.15
Polynomial (Variance of Residuals)	0.9116	0.9510	0.18

Table 5.23.: Differences in area under ROC curves between men and women

5.4. Discussion

The two methods of fitting the non-linear model give identical results. However, fitting the model as two separate linear models can be quicker. This is because nume-

rical differentiation is required for the non-linear method, which is time-consuming. The number of parameters in the model is $n+3p+3$ if a single magnification factor is fitted and $3(n+p+1)$ if three magnification factors are fitted. The time and storage required to fit the model therefore increases rapidly with both the number of subjects used to calculate the model from and the number of parameters in the polynomial part of the model.

The best methods of defining prevalent deformities we have seen are those which look for heights that are lower than might be expected. In general, it seems that looking for a relative loss of height is better than looking for an absolute loss of height. This makes sense since vertebrae differ considerably in size, but less so than in shape. Previous morphometric methods have revolved around identifying vertebrae of an unusual shape, and a relative change in height corresponds to a similar change in shape in different vertebrae, but the same absolute loss of height will change the shape of a large vertebra less than it will change the shape of a smaller vertebra.

Whether the same threshold should be used for all sites and vertebrae or whether site and vertebra specific ones are required is open to debate. If the same threshold is used at all vertebrae, a height that is reduced by, say, 20% will always be classed as a deformity, and hence the sensitivity will be the same for all vertebrae. However, if some vertebrae vary more in shape than others, then those which vary more will be more likely to be incorrectly classed as deformities, so the specificity of the method will differ between vertebrae.

If, on the other hand, a site specific threshold is used, this is effectively ensuring

that the specificity of the method is the same for all vertebrae. However, in this case, a given reduction in height may lay beyond the threshold for one vertebra but not for another, so the sensitivity of the method will differ between vertebrae.

Using the within subject variance of the residuals to determine whether there is a deformity is not dependent on a good fit to be successful. In this population, it did not work as well as the other methods, but multiple fractures were rare in these subjects. In a subject with multiple fractures, the variance of the residuals will increase (there will be a greater proportion of outliers). The methods based on height reduction will be less successful in such a population, since the predicted heights will be biased downwards by the fractures. Even if the model is fitted robustly, if more than half of the vertebral heights are affected, the model will be fitted to the fractured heights, rather than the unfractured heights. Therefore it is possible that in a severely osteoporotic population, such as may be recruited for a clinical trial, the within-subject variance method will prove superior.

The Minne method did not perform as well as the other methods, either in predicting vertebral heights in subjects without deformities, or in identifying subjects with deformities. This can be explained by the fact that only a 3^{rd} order polynomial was used to fit to the heights, and we have seen that a higher order polynomial provided a better fit. In addition, only a single height (T4) is used in fitting the model, which is very inefficient and can be expected to give imprecise estimates of the heights, as it did.

The McCloskey-Kanis method was more precise in predicting the vertebral heights, but was biased, with the heights typically being overestimated by 0.3-0.6mm. This

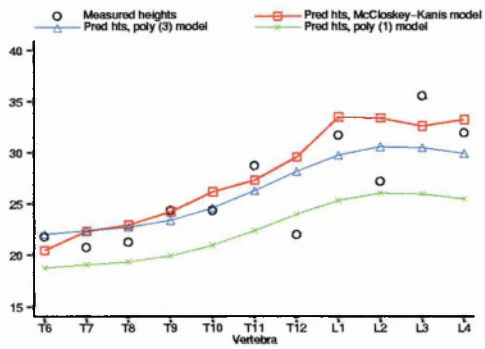
can be explained by the method used to calculated the predicted posterior height.

It was a little disappointing that the McCloskey-Kanis method appeared to be better than the polynomial method at identifying deformed vertebrae. There are two possible reasons for this. One is that using a single magnification factor throughout the spine is inappropriate, since the height measurements are made from two separate films, which may be at different magnifications. Since the McCloskey-Kanis method only uses the 4 adjacent vertebrae to predict heights, it will be less affected by this problem. We can try to avoid it by allowing the magnification to vary along the length of the spine: this is investigated further in chapter 8.

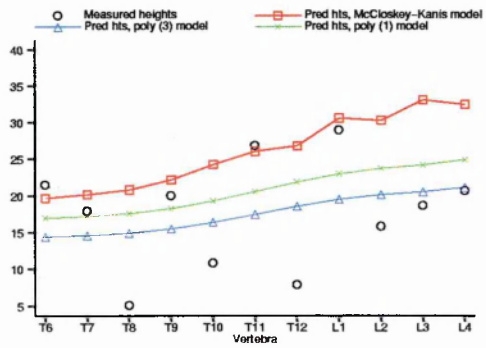
Alternatively, it may be that the polynomial models lack robustness. If a single magnification factor is used for all three sites, all heights are used in predicting each height, whilst the McCloskey-Kanis method excludes heights that are unusually small from the prediction. In a subject with several fractures, the predicted heights from the polynomial model will be biased downwards, due to the heights in the fractures being much smaller than expected. If three separate magnification factors are used, the bias may be greater in one site than another: typically fractures affect the mid height but not the posterior height, so that the mid heights will be underestimated by more than the posterior heights.

This is illustrated in figure 5.4. This subject had fractures at T8, T10, T12 and L2. It can be seen that the predicted heights from the McCloskey-Kanis model are all close to the measured height of the undeformed vertebrae. The three magnification factor model fits the posterior heights reasonably well, but underestimates the unaffected mid heights considerably. The single magnification factor model is less

biased in the mid heights, since the unaffected posterior heights pull it up, but is more biased in the posterior heights since the affected mid heights pull it down.



(a) Posterior Heights



(b) Mid Heights

Figure 5.4.: Measured and predicted posterior and mide vertebral heights in a subject with multiple fractures

We therefore need to develop the polynomial model to avoid using unusually low heights. This is called "robust" estimation. The development and performance of a robust polynomial model are outlined in chapter 6. Since using a single magnification factor seems to be more robust than using three separate ones, this is what was done.

6. Robust Polynomial Models of the Spine

6.1. Introduction

The method of fitting polynomial models introduced in the previous chapter has one very serious drawback: the models need to be defined using a population known to be free from fractures. However, this means that we must use a subjective definition of fracture to define this population, before we can start to use our more objective method. In addition, as we have seen, the predicted heights in subjects with fractures will be biased downwards, and thus genuine fractures will be harder to detect.

Ideally, we would like to have methods of both defining and fitting the polynomial model which will automatically exclude heights of fractured vertebrae from the process. This can be achieved using robust regression. This has the further advantage that the model definition and fitting stages can be combined, since the model no longer needs to be defined on a population free from fractures.

6.2. Robust Regression

6.2.1. Introduction

Regression is a method of predicting a dependent variable from a number of explanatory variables. However, if there are any large errors in the either the predictor variables or the outcome variables, ordinary regression models can give predictions which are a long way from the truth. This can happen in two ways:

1. There are errors in the observations used to define the regression model, either in the dependent or explanatory variables. This can cause the model parameters to be inaccurate, and hence all predictions from the model to be incorrect.

2. Even if the model parameters are correct, if there are errors in the explanatory variables, when they are used to make a prediction, the prediction will also be incorrect.

The first problem is a problem in model definition, the second a problem of model fitting. In the polynomial model, we only need to concern ourselves with the first problem. The explanatory variables in this case are simply the vertebral numbers and powers of them, and so we need not worry about errors in them[1]. Even when defining the model, errors in the predictors can be ignored. This is important, since

[1]It is conceivable that they are incorrect, if the radiographer measuring the vertebral heights makes an error in identifying the vertebral levels, and measured T3 – L3 or T5 – L5 rather than T4 – L4

most methods of robust regression are not robust to x-outliers, and the methods that are are complex and not commonly implemented in standard stats packages.

6.2.2. Robust Model Definition

It is not possible to fit nonlinear models robustly with standard statistical packages. There are, however, widely used methods for fitting linear models robustly, which can easily be adapted for nonlinear models. The most common method was devised by Huber. Using this method, ordinary least squares regression is used as a starting point. The regression is then repeated, but each observation is given a weight, depending on how large the residual is. Observations with large residuals are given smaller weights than observations with small residuals, thus reducing the influence of outlying observations.

It has been suggested [22] that 3 iterations be made using the weights

$$w_i = \begin{cases} 1, & |r_i| \leq ks^* \\ ks^*/|r_i|, & \text{otherwise} \end{cases} \tag{6.1}$$

where $k = 2.4388$, r_i is the residual from the i^{th} observation and $s^* = \text{median}(|r_i|)$. Using these weights, the influence of an observation is reduced if $|r_i|$ exceeds about 1.645 standard deviations (i.e approximately 10% would be trimmed from the tails of a normal distribution). The weights should then be changed to

$$w_i = \begin{cases} (1 - (r_i/cs^*)^2)^2, & (r_i/cs^*)^2 < 1 \\ 0, & \text{otherwise} \end{cases} \tag{6.2}$$

where c is a tuning constant, and iteration continued until the weights w_i do not change by more than 10^{-4} between iterations. If $c = 6$, then zero weight is assigned

if $|r_i| > 4$ standard deviations(SD), whilst if $c = 9$, then zero weight is assigned if $|r_i| > 6$ SD. This is the default procedure implemented in stata as `rreg`.

It is possible to fit equation 5.1 using standard non-linear methods, and from its fit, calculate suitable weights. We saw in Chapter 5 that fitting a single magnification factor to all three sites will be more robust that fitting 3 separate magnification factors. The entire model could then be refitted using these weights, and the procedure repeated until no further change in the model occured. However, this involves fitting the entire model iteratively, and we have already seen that fitting this model even once is time-consuming.

An alternative method is to fit equation 5.1 in two parts, as we did in chapter 5. After fitting both the magnification and polynomial parts of the model, the residuals can be calculated, and from them the weights as defined in equations 6.1 or 6.2. For the next iteration, these weights can be used in calculating the magnification part of the model. However, when calculating the polynomial part of the model, we already need to use weights $1/m_j^2$, so to fit the model robustly we will need to use the weights w_i/m_j^2 for this part of the model.

6.2.3. Robust Model Fitting

If a vertebra is fractured, at least one of the heights in that vertebra will be lower than expected. When fitting a polynomial model, this unusually low height will tend to reduce the overall magnification factor fitted, and hence all the heights will be slightly underestimated. This means that the reduction in height in the fractured vertebra is underestimated, making it slightly harder to identify fractured

vertebrae. Identifying fractured vertebrae may therefore be made easier by using robust methods to fit the model.

As we have seen, once the polynomial model has been defined, fitting the model to a new population does not involve recalculating the a_{jk} parameters from equation 5.1. Thus we are simply fitting a linear model for the m_j, and standard methods of robust regression can be used. The same weighting methods as outlined in the previous section were used. However, since estimating s^* for each new subject separately is likely to be very inefficient, the estimate of s^* derived when the model was originally developed was used in the robust fitting algorithm for all subjects.

6.2.4. Comparing Robustly and Non-Robustly Defined Models

Robust models have the advantage of being less biased if there are outlying observations. However, this is compensated for by the fact that they are less precise than a conventional least squares model if there are no outliers. It is important to quantify the loss of precision by using these robust methods. Therefore, we fitted both robust and least squares models to all of the training sets. We examined the agreement between the predicted heights from each of the models, and also the standard deviation of the residuals. We would expect that for the observations given large weights in the robust regression, the fit would be better than for the least squares model, whilst the reverse would be true for the observations given small weights.

We also need to see how robust the models are to fractures. We therefore took the training set and randomly selected 5% of heights to be reduced by varying amounts up to 10mm. The reductions followed a uniform distribution. This will

cover very slight deformities that we would not expect to be able to detect up to very large deformities that any method should detect. A robust polynomial model was fitted to the new data, and the standard deviation of the residuals compared to the standard deviation of the residuals from the least squares model on the unaltered data, restricted to the unaltered heights. This gives a measure of the loss of precision caused by the height reductions.

The robustly defined model was fitted to all subjects using both robust and non-robust methods. The mean and standard deviation of the residuals in subjects without fractures were calculated using both methods. However, the most interesting comparison of the residuals is between fractured and unfractured vertebrae in subjects with fractures. We can expect the residuals to be biased downwards in the unfractured vertebrae in these subjects using non-robust fitting, but not using robust fitting.

The largest reduction in height and largest relative reduction in height were taken as test statistics. As in chapter 5, two thresholds were used to define fractures, one to give the same sensitivity as the McCloskey-Kanis method and one to give the same specificity. The abilities to distinguish both fractured vertebrae and subjects with fractured vertebrae were compared using ROC curves.

6.3. Results

6.3.1. Effect of Robust Model Definition

Effect in Subjects Without Fractures

The differences between the robustly and non-robustly defined models were extremely small. The correlation between the predictions from the two models was greater than 0.99, with the differences lying in the range -0.2mm to 0.3mm. The means and standard deviations of the residuals from both models are given in Table 6.1. It can be seen that the differences between the two models are slight: there has not been a great loss in efficiency by using robust methods.

Effect in Subjects with Simulated "Fractures"

The means and standard deviations of the residuals from the robustly and non-robustly defined models in the population of men from Heidelberg with 5% of the heights artificially reduced is shown in table 6.2. The mean and standard deviation of the residuals from the polynomial model with no heights reduced is also given for comparison, in those heights that were not reduced. Only those heights which were not "fractured" are included.

Clearly, the non-robust method leads to a slight underestimation of all heights if some vertebrae are fractured. However, using the robust method is as accurate and precise as when there were no fractures in the population in which the model was defined.

		Training Set		Other Normals	
		Mean	SD	Mean	SD
Heidelberg Men	Robust	-0.000	1.491	-0.010	1.454
	Non-Robust	-0.001	1.490	-0.011	1.454
Heidelberg Women	Robust	-0.000	1.238	0.014	1.224
	Non-Robust	-0.000	1.238	0.014	1.226
Malmo Men	Robust	0.001	1.501	0.011	1.582
	Non-Robust	0.000	1.499	0.010	1.577
Malmo Women	Robust	-0.001	1.360	-0.032	1.457
	Non-Robust	-0.001	1.358	-0.031	1.460
Graz Men	Robust	0.000	1.204	-0.001	1.322
	Non-Robust	0.000	1.203	-0.001	1.322
Graz Women	Robust	0.005	1.218	0.027	1.210
	Non-Robust	0.001	1.217	0.023	1.207

Table 6.1.: Residuals from robust and non-robust model definition

Population	Model	Mean	SD
Fractures	Non-Robust	0.283	1.482
Fractures	Robust	-0.016	1.457
No Fractures	Robust	0.014	1.486

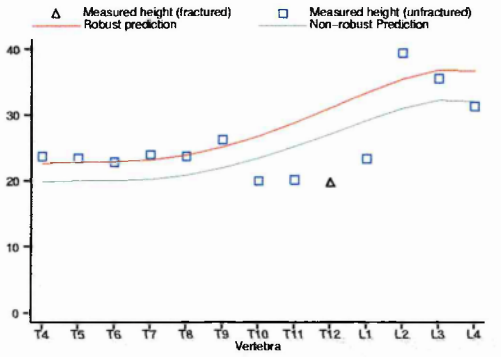
Table 6.2.: Means & standard deviations from models defined in a population with simulated fractures

6.3.2. Effect of Robust Model Fitting

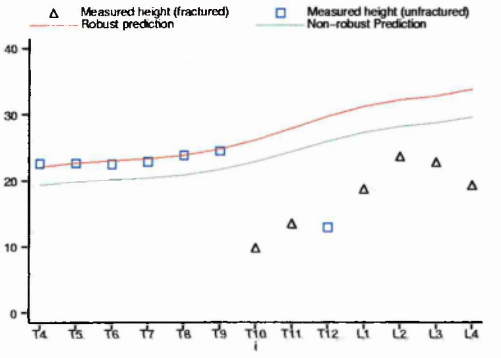
The effect of fitting the model robustly is illustrated in figure 6.1. This shows the vertebral heights of a single individual with deformities at T10 to L4. The measured heights judged by the radiologist to be unaffected by fracture are shown as squares, whilst those judged to have been reduced by a fracture are shown as triangles. None of the posterior heights are affected, but 6 mid-heights and one anterior height are. It is also clear that several other anterior heights are lower than may have been expected, despite the deformities being classed as concave or biconcave.

The line with plus signs gives the predicted heights from a non-robust fitting of the polynomial model, whilst the line with open circles gives the predicted heights from the robust fitting. It can be seen that the non-robustly predicted heights are uniformly less than the robustly predicted heights. This is because the fractured heights are very much less than expected and since all heights are given equal weight in the model fitting, the line is pulled downwards. When the model is fitted robustly, these heights are regarded as outliers and given a reduced weighting when fitting the model, so the fit to the unfractured heights is better.

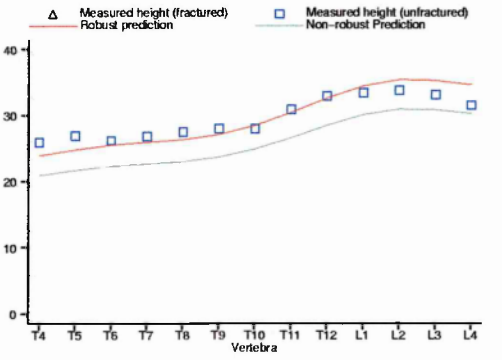
This has two effects. One is to improve the prediction of the unfractured heights, both in terms of accuracy (they are no longer consistently underestimated) and precision (the magnitude of the residuals in the unfractured heights are smaller). Secondly, the magnitude of the residuals of the fractured heights is increased, since the fitted line is now closer to the unfractured heights and further from the fractured heights. Both of these effects will tend to increase the difference between the



(a) Anterior Heights



(b) Mid Heights



(c) Posterior Heights

Figure 6.1.: Measured and predicted heights in one subject with multiple fractures

residuals from fractured and unfractured heights, making it easier to distinguish between them.

Table 6.3 shows the effect of using robust methods to fit the model. The mean and SD of the residuals, using both robust and least squared fitting, are given for the following groups heights:

1. All heights in the training set
2. All heights in the testing set
3. All heights in fractured vertebrae
4. All heights in unfractured vertebrae in subjects with fractures

Heights	Least Squares Fit		Robust Fit	
	Mean	SD	Mean	SD
Training set	0.00	1.34	-0.02	1.35
Testing set	0.00	1.38	-0.03	1.38
Fractured vertebrae	-3.28	4.71	-3.86	4.88
Unfractured vertebrae	0.45	1.70	0.06	1.66

Table 6.3.: Comparison of robust and least squared fitting on accuracy and precision of prediction

The differences between the two methods of fitting are minimal in the training and testing sets. However, in subjects with fractures, the unfractured vertebrae are estimated more accurately and precisely using the robust fitting method than

	Sensitivity	Specificity
Fixed threshold	15.3%	16.1%
Vertebra specific threshold	2.94 SD	3.32 SD

Table 6.4.: Thresholds used to define fractures in robust polynomial models

using the least squares method. The difference between the predicted and observed heights in the fractured vertebrae is greater using the robust fitting method, so we can hope that this method will perform better at identifying fractures.

6.3.3. Identification of Deformed Vertebrae using the Robust Polynomial Model

The thresholds chosen to define fractures are given in table 6.4.

The vertebra specific thresholds varied from 13.6% to 22.8% to give the same specificity as the McCloskey-Kanis model, and 12.2% to 20.5% to give the same sensitivity. We can already see from Table 6.4 that this model does not perform as well as the McCloskey-Kanis model, since the thresholds to give the same specificity are more extreme than those to give the same sensitivity. Thus, this model will be less sensitive at the same level of specificity and less specific at the same level of sensitivity. However, these thresholds are more extreme than those used with the least squares polynomial model, suggesting that this robust model should perform better than the least squares model.

The actual numbers of vertebrae classed as deformities in this way is given in Tables 6.5 and 6.6.

Population	Site-Specific	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	No	13/910=1.4%	5/701=0.7%	7/271=2.6%	20/35=57.1%	12/476 =2.5%
Heidelberg Men	Yes	10/910=1.1%	4/701=0.6%	4/271=1.5%	17/35=48.5%	6/476=1.3%
Heidelberg Women	No	3/910=0.3%	2/811=0.3%	6/208=2.4%	45/49=91.8%	3/139=1.9%
Heidelberg Women	Yes	4/910=0.4%	5/811=0.6%	6/208=2.9%	45/49=91.8%	6/139=4.3%
Malmo Men	No	5/910=0.6%	10/1022=1.0%	9/439=2.1%	48/61=78.7%	20/209=9.6%
Malmo Men	Yes	5/910=0.6%	12/1022=1.2%	9/439=2.1%	44/61=72.1%	18/209=8.6%
Malmo Women	No	6/910=0.7%	11/1218=0.9%	11/450=2.4%	54/62=87.1%	8/114=7.0%
Malmo Women	Yes	6/910=0.7%	18/1218=1.5%	12/450=2.7%	54/62=87.1%	9/114=7.9%
Graz Men	No	3/910=0.3%	11/1025=1.1%	3/389=0.8%	45/54=83.3%	6/241=2.5%
Graz Men	Yes	4/910=0.4%	11/1025=1.1%	7/389=1.8%	42/54=77.8%	7/241=2.9%
Graz Women	No	4/910=0.4%	7/1391=0.5%	1/225=0.4%	20/30=66.7%	1/104=1.0%
Graz Women	Yes	5/910=0.6%	9/1391=0.7%	1/225=0.4%	20/30=66.7%	2/104=1.9%
Total	No	34/5460 = 0.6%	46/6168 = 0.8%	37/1982=1.9%	232/291=79.7%	50/1283=3.9%
Total	Yes	34/5460 = 0.7%	72/6168 = 1.0%	39/1982=2.0%	222/298=75.5%	48/1283=3.7%

Table 6.5.: Number of deformities from robust polynomial model with same specificity as McCloskey-Kanis model in training samples

Population	Site-Specific	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	No	17/910=1.9%	6/701=0.9%	6/271=3.0%	20/35=57.1%	14/476=2.9%
Heidelberg Men	Yes	16/910=1.8%	7/701=1.0%	5/271=1.9%	21/35=60.0%	12/476=2.5%
Heidelberg Women	No	5/910=0.6%	4/811=0.5%	6/208=2.9%	45/49=91.8%	4/139=2.9%
Heidelberg Women	Yes	11/910=1.2%	8/811=1.0%	9/208=4.3%	46/49=93.9%	10/139=7.2%
Malmo Men	No	7/910=0.8%	15/1022=1.5%	11/439=2.5%	52/61=85.3%	22/209=10.5%
Malmo Men	Yes	6/910=0.7%	23/1022=2.3%	14/439=3.2%	54/61=88.5%	22/209=10.5%
Malmo Women	No	10/910=1.1%	17/1218=1.4%	14/450=3.1%	54/62=87.1%	9/114=7.9%
Malmo Women	Yes	11/910=1.2%	33/1218=2.7%	16/450=3.6%	54/62=87.1%	12/114=10.5%
Graz Men	No	4/910=0.4%	14/1025=1.4%	7/389=1.8%	48/54=88.9%	7/241=2.9%
Graz Men	Yes	9/910=1.0%	27/1025=2.6%	12/389=3.1%	45/54=83.6%	10/241=4.2%
Graz Women	No	9/910=1.0%	12/1391=0.9%	1/225=0.4%	21/30=70.0%	2/104=1.9%
Graz Women	Yes	10/910=1.1%	18/1391=1.3%	2/225=0.9%	20/30=66.7%	3/104=2.9%
Total	No	52/5460 = 1.0%	68/6168 = 1.1%	47/1982=2.4%	240/291=82.5%	58/1283=4.5%
Total	Yes	63/5460 = 1.2%	116/6168 = 1.9%	58/1982=2.9%	240/291=82.5%	69/1283=5.0%

Table 6.6.: Number of deformities from polynomial model with same sensitivity as McCloskey-Kanis model in training samples

Again, the polynomial model performs slightly less well than the McCloskey-Kanis model. With the same number of false positives in the training sample (34), the polynomial model detects slightly fewer fractures (223 or 232 vs 240), whilst with the same number of true positives (240) there are slightly more false positive with the polynomial model rather than the McCloskey-Kanis model (52 or 63 vs 34). Using the same threshold for all heights seems to work better than site-specific thresholds.

6.3.4. Identification of Subjects with Deformities Using the Robust Polynomial Model

Table 6.3.4 shows the classification of subjects as cases or non-cases using the robust polynomial model, according to whether or not they had vertebral fractures on the clinical reading of the films. The values in this table differ only very slightly from those in table 5.21: the added complexity of robustly fitting the polynomial model has made very little difference to its ability to differentiate between subjects with fractures and those without.

6.4. Discussion

6.4.1. Robust Model Definition

We have seen that the losses of precision and accuracy caused by using robust methods to define our model in the absence of fractured vertebrae are small. If there are fractures in the population used to define the model, there is a marked loss of

Cutoff	Training Set	Testing Set	Fractures	Other Deformities
15%	47/420=11.2%	57/489 = 11.7%	158/178 = 88.8%	40/105 = 38.1%
15.5%	45/420=10.7%	50/489 = 10.2%	152/178 = 85.4%	40/105 = 38.1%
16.3%	24/420= 5.7%	30/489 = 6.1%	145/178 = 81.5%	35/105 = 33.3%
20%	11/420= 2.6%	3/489 = 0.6%	126/178 = 70.8%	22/105 = 20.1%

Table 6.7.: Cross-tabulation of morphometric and clinical classifications of the presence of at least one deformity in a subject

accuracy using non-robust methods, but again the robust methods are accurate and precise.

By using robust methods to define our model, it is not necessary to define the model on subjects known to be free from fractures. This is a great advantage since it makes the method "free-standing": there is no longer any need for a radiologists opinion at any stage in detecting deformities, and hence the method is completely objective, as are the other methods in common use.

The method of robust regression used is sensitive to outliers in the predictor variables, but not outliers in the outcome variable. Since we are using polynomial models, where the predictors are simply powers of the vertebral levels, it is not possible to have outliers in the predictor variables. Thus this form of robust regression is perfectly adequate. Methods that are insensitive to both x and y-outliers are generally far more computationally intensive, and the addition effort is not justified

in this case.

Only 5% of heights were artificially fractured. It is possible that if the proportion of fractured heights were greater than this, the method would break down. However, within the EVOS population as a whole, only 15% of subjects had a deformity and less than 2% of vertebrae were fractured, and probably far less than 2% of heights. In a clinical trial setting, there may be far more fractures (since the presence of at least one vertebral deformity is a common inclusion criterion), so the method may be less successful.

6.4.2. Robust Model Fitting

Again, we do not need to concern ourselves with x-outliers in this context, so the low breakdown point to such outliers of the regression method we are using is not a problem.

Breakdown due to y-outliers may, however, be a problem. Although the form of robust regression we are using here has a comparatively low breakdown point $(1/p + 1)$, we are only fitting a single variable (the magnification factor), so we can expect a breakdown point of 50%. I.e. we can expect the method to work reasonably as long as less than half of the heights are affected by fractures. Since only a small proportion of fractures affect the posterior height, even if half of the vertebrae are affected, less than half of the heights are likely to be. However, this is a "hard" limit on any robustness algorithm: if there is more bad data than good data, it becomes impossible to identify the good data.

The effects shown in figure 6.1 are fairly small: the difference between the robust

and non-robust predicted heights being between 2.5 and 4.5mm. Furthermore, this subject was chosen because she had multiple deformities, which makes the effect easier to see. It is by no means obvious that this effect will be of great importance in general, since multiple deformities are comparatively rare.

7. Latent Variable Models of the Spine

7.1. Introduction

An alternative way to think of the patterns in the vertebral height measurements is in terms of a latent variable model. In such a model, the manifest variables (heights) are thought of as a linear combination of a smaller number of latent factors. We can use the measured heights to obtain estimates of the values of the latent variables, then use these latent variable values to obtain predicted heights.

7.2. Factor Analysis Model

Suppose the 39 observed heights in the i^{th} subject are $x_{i1}, x_{i2}, \ldots, x_{i39}$. The factor analysis model suggests that each variable x_{ij} is a linear function of a set of k latent variables $f_{i1}, f_{i2}, \ldots, f_{ik}$, where $k \leq 39$, plus a residual term u_{ij}. It is usually convenient in factor analysis to have the mean of each variable equal to 0, so we

calculate

$$\mathbf{z_i} = \mathbf{x_i} - \mathbf{M}$$

where $\mathbf{M}$ is a vector of the means of the variables $x_{.1} \ldots x_{.39}$. Then the factor analysis model is

$$\mathbf{z_i} = \mathbf{\Lambda f_i} + \mathbf{u_i}$$

where $\mathbf{f_i'} = [f_{i1}, f_{i2}, \ldots, f_{ik}]$, $\mathbf{u_i'} = [u_{i1}, u_{i2}, \ldots, u_{i39}]$ and $\mathbf{\Lambda}$ is a $39 \times k$ matrix containing the *factor loadings.*

Conventional factor analysis gives Λ (in practice, iterated principal factors were calculated by using the `ipf` option in the stata `factor` routine), but calculating the vector of factor scores, f_i is more complicated. If k is less than 39, then there are a number of different solutions for f_i. There are several methods of calculating them: the simplest is to treat the problem as a regression. The regression equation can be written as

$$\mathbf{F} = \mathbf{ZW} + \mathbf{E} \tag{7.1}$$

where F is the $n \times k$ matrix of factor scores, Z is the $n \times p$ matrix of centred variables z W are the regression coefficients used to calculate F from Z, and E is an $n \times k$ matrix of error terms. It is not possible to solve this by regression, since we do not have an estimate of F. However, the regression solution for this equation is

$$\mathbf{W} = (\mathbf{Z'Z})^{-1}(\mathbf{Z'F})$$

and the term $(\mathbf{Z'Z})^{-1}$ is straightforward to calculate (it is n times the covariance matrix of the centred variables). The term $(\mathbf{Z'F})$ cannot be calculated directly, since $\mathbf{F}$ is unknown. However, $(\mathbf{Z'F})/\mathbf{n}$ is the covariance matrix between the variables and

the factors, and this matrix is produced in the course of the factor analysis. Thus $\mathbf{W}$ can be calculated, and from $\mathbf{W}$, $\mathbf{F}$.

Having calculated the factor scores, we can obtain predicted values for the the z_{ij}, given by

$$\hat{\mathbf{z}}_\mathbf{i} = \Lambda \mathbf{f}_\mathbf{i}.$$

and hence

$$\hat{\mathbf{x}}_\mathbf{i} = \Lambda \mathbf{f}_\mathbf{i} + \mathbf{M}$$

7.2.1. Identifying Deformities

If a fracture has occurred, the fractured height will be less than its predicted value. This fact can be used to identify deformities. The difference between the expected value and the measured value of x_{ij} is u_{ij}. We can use this value directly to define fracture: if u_{ij} is less than a particular value the vertebra is classed as fractured. For example, we may use the condition $u_{ij} \leq -4$: i.e. a vertebra is classed as a fracture if the measured height is 4mm or more less than the expected height. Alternatively we may consider a relative reduction, by using the test statistic $u_{ij}/\hat{x_{ij}}$. Since we have seen with the polynomial models that relative reductions tend to work better than absolute reductions when defining fractures, we will use the relative reductions.

7.3. Results

7.3.1. Numbers of Factors

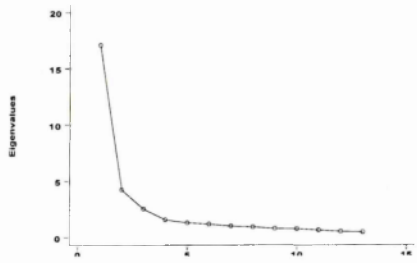
Factors were derived from the covariance matrices of each of the training sets using iterated principal factors. Figure 7.1 shows plots of the size of the 13 largest eigenvalues of each matrix. A commonly used method of deciding how many factors to retain is to look for an "elbow" in these plots, where the slope becomes markedly less steep. Only factors to the left of the elbow explain considerably more variation the factors to their right, and thus should be retained.

It can be seen that in some cases, only one factor would be considered as important, whilst in other cases up to 3 could be. Since we want to use the same model for all populations, we will consider the first 3 factors.

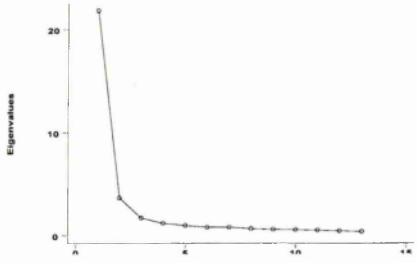
7.3.2. Interpretation of Factors

First Factor

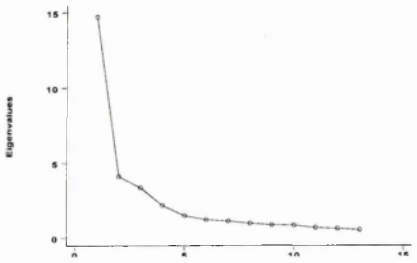
Table 7.1 shows the coefficients of the first factor in each of the 6 training populations. It can be seen that in all six groups, the coefficients of each measurement are similar. Thus this first measurement is akin to a magnification factor (as this factor increases, every height in the spine increases, albeit not by exactly the same proportion).



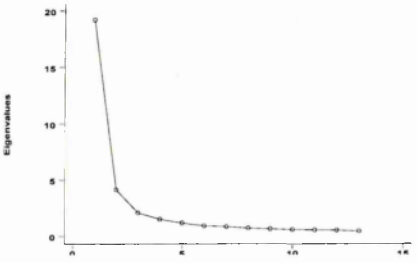
(a) Men from Heidelberg



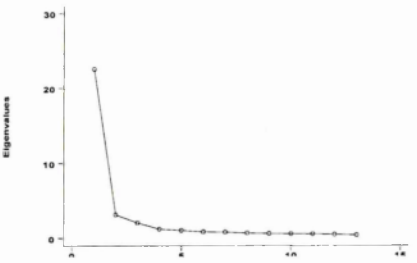
(b) Women from Heidelberg



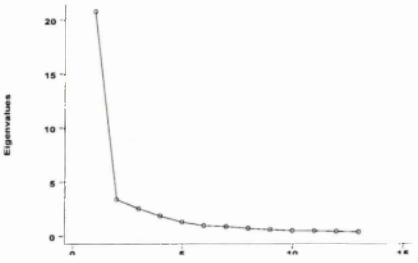
(c) Men from Malmo



(d) Women from Malmo



(e) Men from Graz



(f) Women from Graz

Figure 7.1.: 'Scree' plots of eigenvalues from covariance matrices of each training set

Height	Heidelberg		Malmo		Graz	
	Men	Women	Men	Women	Men	Women
T4 ant.	0.150	0.146	0.123	0.149	0.127	0.096
T4 mid.	0.164	0.163	0.095	0.170	0.153	0.139
T4 post.	0.158	0.138	0.102	0.157	0.159	0.161
T5 ant.	0.172	0.165	0.158	0.145	0.130	0.145
T5 mid.	0.160	0.175	0.138	0.139	0.161	0.159
T5 post.	0.124	0.163	0.141	0.161	0.166	0.172
T6 ant.	0.180	0.145	0.141	0.174	0.141	0.125
T6 mid.	0.191	0.163	0.177	0.177	0.163	0.167
T6 post.	0.171	0.164	0.161	0.170	0.164	0.169
T7 ant.	0.181	0.137	0.174	0.183	0.144	0.154
T7 mid.	0.195	0.162	0.174	0.183	0.164	0.182
T7 post.	0.186	0.176	0.152	0.177	0.177	0.175
T8 ant.	0.164	0.139	0.168	0.185	0.156	0.164
T8 mid.	0.164	0.167	0.188	0.179	0.176	0.177
T8 post.	0.178	0.185	0.185	0.177	0.174	0.186
T9 ant.	0.153	0.141	0.177	0.151	0.173	0.149
T9 mid.	0.184	0.182	0.172	0.193	0.169	0.182
T9 post.	0.187	0.178	0.180	0.178	0.170	0.190
T10 ant.	0.147	0.158	0.158	0.155	0.160	0.144
T10 mid.	0.177	0.187	0.161	0.194	0.167	0.169
T10 post.	0.177	0.177	0.186	0.171	0.145	0.180
T11 ant.	0.127	0.137	0.176	0.140	0.155	0.146
T11 mid.	0.160	0.182	0.191	0.177	0.163	0.166
T11 post.	0.142	0.188	0.180	0.169	0.156	0.161
T12 ant.	0.134	0.156	0.167	0.136	0.154	0.149
T12 mid.	0.175	0.170	0.166	0.151	0.178	0.169
T12 post.	0.176	0.171	0.130	0.143	0.179	0.171
L1 ant.	0.123	0.156	0.165	0.155	0.156	0.166
L1 mid.	0.147	0.171	0.178	0.149	0.170	0.166
L1 post.	0.157	0.163	0.182	0.147	0.183	0.169
L2 ant.	0.140	0.151	0.176	0.122	0.144	0.145
L2 mid.	0.159	0.157	0.156	0.148	0.156	0.169
L2 post.	0.175	0.155	0.182	0.136	0.169	0.162
L3 ant.	0.140	0.113	0.139	0.142	0.149	0.141
L3 mid.	0.156	0.153	0.147	0.133	0.164	0.159
L3 post.	0.134	0.155	0.165	0.160	0.159	0.155
L4 ant.	0.126	0.145	0.132	0.145	0.144	0.135
L4 mid.	0.123	0.149	0.112	0.149	0.153	0.159
L4 post.	0.137	0.128	0.120	0.136	0.156	0.134

Table 7.1.: Coefficients of unit vectors of first factor in each population

Second Factor

The second factor has positive loadings on the heights in the thoracic spine but negative loadings on the height in the lumbar spine. We have seen that heights tend to increase from the thoracic to the lumbar spine, and this factor is a measure of the magnitude of this increase. Thus this factor measures differences in shape, rather than differences in size, of the spine.

Third Factor

The third factor does not have a consistent interpretation across centres, although there is a suggestion of a quadratic pattern in some groups (positive coefficients in the upper and lower spine, negative coefficients in the middle of the spine). This lack of pattern is not surprising, since in some centres there appeared to be only one or two genuine factors, so the third factor will represent random noise in these centres.

7.3.3. Correlations between Factor Scores

Using different training sets to define the factors leads to different factors. Tables 7.4, 7.5 and 7.6 show the correlations between the three factor scores defined using the six different populations. Clearly, it does not matter which population is used to define the first factor, since the correlations are all above 0.99. There is some disagreement in the second factor and even more in the third factor. However, all of the correlations are very highly significant ($p < 0.0002$).

Note that the sign of a factor score is arbitrary, so a negative correlation between

Height	Heidelberg		Malmo		Graz	
	Men	Women	Men	Women	Men	Women
T4 ant.	-0.192	-0.177	-0.096	-0.199	0.202	-0.113
T4 mid.	-0.157	-0.116	-0.014	-0.160	0.167	-0.143
T4 post.	-0.095	-0.090	-0.018	-0.161	0.126	-0.169
T5 ant.	-0.190	-0.227	-0.160	-0.240	0.309	0.084
T5 mid.	-0.174	-0.194	-0.057	-0.267	0.237	-0.046
T5 post.	-0.105	-0.193	-0.109	-0.232	0.206	-0.067
T6 ant.	-0.161	-0.209	-0.158	-0.190	0.220	0.076
T6 mid.	-0.144	-0.209	-0.107	-0.138	0.221	-0.047
T6 post.	-0.102	-0.184	-0.137	-0.221	0.189	-0.040
T7 ant.	-0.154	-0.217	-0.111	-0.098	0.234	0.043
T7 mid.	-0.142	-0.171	-0.098	-0.134	0.205	-0.027
T7 post.	-0.063	-0.111	-0.182	-0.175	0.133	-0.046
T8 ant.	-0.180	-0.168	-0.164	-0.061	0.204	-0.007
T8 mid.	-0.205	-0.188	-0.133	-0.118	0.112	-0.088
T8 post.	-0.085	-0.096	-0.215	-0.168	0.039	-0.020
T9 ant.	-0.162	-0.118	-0.155	0.021	0.041	-0.136
T9 mid.	-0.079	-0.096	-0.182	-0.001	0.016	-0.120
T9 post.	0.072	-0.092	-0.247	-0.060	-0.101	-0.080
T10 ant.	-0.139	0.072	-0.050	0.094	-0.070	-0.161
T10 mid.	-0.083	0.046	-0.139	-0.047	-0.122	-0.218
T10 post.	-0.034	0.040	-0.069	0.076	-0.163	-0.131
T11 ant.	0.059	0.201	0.050	0.063	-0.120	-0.115
T11 mid.	0.025	0.119	0.039	0.097	-0.158	-0.208
T11 post.	0.081	0.032	-0.030	0.135	-0.178	-0.192
T12 ant.	0.144	0.197	0.105	0.101	-0.151	-0.137
T12 mid.	0.129	0.167	0.061	0.128	-0.127	-0.209
T12 post.	0.096	0.044	0.030	0.157	-0.118	-0.131
L1 ant.	0.253	0.210	0.188	0.097	-0.153	0.136
L1 mid.	0.192	0.167	0.252	0.140	-0.149	0.156
L1 post.	0.068	0.075	0.107	0.197	-0.086	0.119
L2 ant.	0.263	0.232	0.100	0.247	-0.216	0.230
L2 mid.	0.254	0.153	0.238	0.176	-0.176	0.202
L2 post.	0.154	0.191	0.223	0.224	-0.157	0.219
L3 ant.	0.272	0.180	0.240	0.244	-0.157	0.245
L3 mid.	0.244	0.132	0.287	0.215	-0.169	0.259
L3 post.	0.254	0.212	0.220	0.193	-0.093	0.241
L4 ant.	0.140	0.162	0.194	0.136	-0.071	0.259
L4 mid.	0.182	0.145	0.279	0.145	-0.088	0.229
L4 post.	0.174	0.198	0.229	0.149	0.068	0.312

Table 7.2.: Coefficients of unit vectors of second factor in each population

Height	Heidelberg		Malmo		Graz	
	Men	Women	Men	Women	Men	Women
T4 ant.	0.265	-0.249	0.342	0.184	0.015	0.283
T4 mid.	0.190	-0.238	0.380	0.089	0.025	0.191
T4 post.	0.229	-0.243	0.288	0.127	-0.085	0.160
T5 ant.	0.226	-0.153	0.218	0.184	0.030	0.254
T5 mid.	0.297	-0.177	0.196	0.156	0.029	0.244
T5 post.	0.399	-0.180	0.234	0.192	0.077	0.209
T6 ant.	0.032	-0.025	0.104	0.037	0.029	0.312
T6 mid.	0.057	-0.038	0.202	0.141	-0.006	0.235
T6 post.	0.189	-0.030	0.125	0.008	-0.025	0.206
T7 ant.	-0.173	0.232	-0.002	-0.098	0.028	0.151
T7 mid.	-0.070	0.174	-0.039	-0.073	-0.038	0.089
T7 post.	-0.001	-0.018	-0.011	-0.013	-0.019	0.136
T8 ant.	-0.269	0.411	-0.100	-0.158	-0.055	-0.047
T8 mid.	-0.155	0.227	0.014	-0.122	-0.025	0.057
T8 post.	-0.021	0.033	0.053	-0.099	-0.091	0.029
T9 ant.	-0.239	0.376	-0.167	-0.264	-0.186	-0.046
T9 mid.	-0.195	0.088	-0.173	-0.122	-0.115	-0.004
T9 post.	-0.018	-0.049	0.015	-0.138	-0.163	0.068
T10 ant.	-0.144	0.221	-0.243	-0.236	-0.271	-0.192
T10 mid.	-0.123	0.016	-0.271	-0.123	-0.272	-0.155
T10 post.	-0.052	-0.080	-0.140	-0.216	-0.299	-0.057
T11 ant.	-0.232	0.264	-0.132	-0.197	-0.119	-0.231
T11 mid.	-0.185	0.117	-0.163	-0.176	-0.173	-0.142
T11 post.	-0.172	-0.098	-0.066	-0.209	-0.200	-0.153
T12 ant.	-0.148	0.050	-0.041	-0.084	-0.138	-0.286
T12 mid.	-0.115	-0.051	-0.131	0.003	-0.083	-0.190
T12 post.	-0.042	-0.076	-0.174	-0.096	-0.086	-0.121
L1 ant.	0.026	0.000	0.075	0.063	0.021	-0.146
L1 mid.	0.013	-0.119	0.037	0.114	0.154	-0.098
L1 post.	0.009	-0.017	-0.128	-0.124	0.119	-0.133
L2 ant.	0.025	-0.076	-0.002	0.083	0.154	-0.012
L2 mid.	-0.010	-0.147	-0.121	0.132	0.192	-0.099
L2 post.	-0.036	-0.044	-0.147	-0.037	0.218	-0.153
L3 ant.	0.091	0.134	0.168	0.149	0.094	0.087
L3 mid.	0.115	-0.103	0.045	0.280	0.146	-0.083
L3 post.	0.100	-0.071	0.030	0.204	0.350	-0.133
L4 ant.	-0.008	0.173	0.146	0.131	0.257	0.101
L4 mid.	0.179	-0.077	0.060	0.304	0.298	-0.090
L4 post.	0.130	0.038	0.102	0.317	0.293	-0.081

Table 7.3.: Coefficients of unit vectors of third factor in each population

	Heidelberg		Malmo		Graz	
	Men	Women	Men	Women	Men	Women
H'berg Men	1.0000					
H'berg Women	0.9995	1.0000				
Malmo Men	0.9989	0.9990	1.0000			
Malmo Women	0.9986	0.9989	0.9985	1.0000		
Graz Men	0.9991	0.9994	0.9989	0.9989	1.0000	
Graz Women	0.9984	0.9986	0.9981	0.9984	0.9986	1.0000

Table 7.4.: Correlations between first factor scores defined in each of the different populations

	Heidelberg		Malmo		Graz	
	Men	Women	Men	Women	Men	Women
H'berg Men	1.0000					
H'berg Women	0.9122	1.0000				
Malmo Men	0.8685	0.9271	1.0000			
Malmo Women	0.9235	0.9490	0.8647	1.0000		
Graz Men	-0.7755	-0.9281	-0.8256	-0.8897	1.0000	
Graz Women	0.5741	0.6446	0.8182	0.5617	-0.5798	1.0000

Table 7.5.: Correlations between second factor scores defined in each of the different populations

	Heidelberg		Malmo		Graz	
	Men	Women	Men	Women	Men	Women
H'berg Men	1.0000					
H'berg Women	-0.7964	1.0000				
Malmo Men	0.7471	-0.4297	1.0000			
Malmo Women	0.8094	-0.5995	0.5654	1.0000		
Graz Men	0.4978	-0.3574	0.1806	0.8305	1.0000	
Graz Women	0.5066	-0.1913	0.8808	0.1961	-0.1785	1.0000

Table 7.6.: Correlations between third factor scores defined in each of the different populations

two factor scores is equivalent to a positive correlation between them of the same magnitude.

Given the strength of the correlations between the factor scores defined using different populations, it seems reasonable to define a single set of scores to use with all populations. The obvious way to do this is to factor the covariance matrix of all six training sets combined. Table 7.7 shows the correlations between the scores defined using separate populations and the scores defined using the entire population.

7.3.4. Distribution of Factor Scores

There were highly significant differences between centres and between genders in the distributions of the factor scores. The mean and standard deviation of each

Population	Factor 1	Factor 2	Factor 3
Heidelberg Men	0.9994	0.9684	0.9035
Heidelberg Women	0.9997	0.9597	-0.5746
Malmo Men	0.9992	0.9307	0.7670
Malmo Women	0.9992	0.9600	0.8977
Graz Men	0.9996	-0.8471	0.6604
Graz Women	0.9987	0.6817	0.5184

Table 7.7.: Correlations between overall factor scores and centre-specific factor scores

score (using the scores defined by analysing all six training sets combined) in each population is given in table 7.8, along with the significance of differences between centres and between sexes assessed using two-way ANOVA.

The first factor was consistently greater in men than in women, which is to be expected since it is a measure of size. It also differed significantly between centres: this may be due to genuine size differences in the populations, or differences in the magnification of the x-rays. The second factor was consistently lower in men than in women, suggesting that the normal shape of the spine differs between men and women. However, the third factor did not differ significantly either between centres or between sexes.

Population	Factor 1	Factor 2	Factor 3
Heidelberg Men	0.391 (0.591)	-0.450 (1.031)	0.024 (1.019)
Heidelberg Women	-0.364 (0.630)	0.428 (0.884)	-0.046 (0.751)
Malmo Men	1.333 (0.524)	-0.151 (0.976)	-0.054 (1.168)
Malmo Women	0.348 (0.605)	0.356 (0.941)	0.162 (0.951)
Graz Men	-0.512 (0.638)	-0.473 (0.796)	0.017 (0.903)
Graz Women	-1.197 (0.576)	0.289 (0.828)	-0.103 (0.973)
Between Sex Differences	$p = 0.0000$	$p = 0.0000$	$p = 0.93$
Between Centre Differences	$p = 0.0000$	$p = 0.25$	$p = 0.69$

Table 7.8.: Distribution of factor scores in training sets

7.3.5. Fit of Model To Training Sets

Table 7.9 gives the mean and standard deviation of the residuals in each of the training sets. If the overall factor scores are used rather than the set specific ones, the standard deviation increases slightly and the mean is no longer 0, but the changes are small, as can be seen in table 7.10.

7.3.6. Fit of Models To Other Normal Subjects

The mean and S.D. of the residuals from fitting the latent variable models to various testing sets is given in table 7.11. The means are again close to 0, albeit less close than in the training sets. The S.D.s are similar to those in the training sets, but slightly larger in most cases.

Again, using the overall model, the fit of the model is similar. In fact, the stan-

	One Latent Variable		Two Latent Variables		Three Latent Variables	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Heidelberg Men	0.000	1.470	0.000	1.310	0.000	1.211
Heidelberg Women	0.000	1.221	0.000	1.081	0.000	1.013
Malmo Men	0.000	1.486	0.000	1.335	0.000	1.236
Malmo Women	0.000	1.347	0.000	1.197	0.000	1.110
Graz Men	0.000	1.193	0.000	1.085	0.000	0.984
Graz Women	0.000	1.219	0.000	1.061	0.000	0.977

Table 7.9.: Mean and standard deviation of residuals from set-specific latent variable model in training sets

	One Latent Variable		Two Latent Variables		Three Latent Variables	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Heidelberg Men	-0.023	1.519	0.005	1.351	0.005	1.273
Heidelberg Women	0.036	1.312	0.008	1.153	0.009	1.104
Malmo Men	-0.015	1.514	-0.005	1.379	-0.005	1.276
Malmo Women	0.019	1.392	-0.004	1.239	-0.005	1.156
Graz Men	-0.030	1.287	-0.000	1.147	-0.000	1.068
Graz Women	0.014	1.263	-0.004	1.133	-0.004	1.035

Table 7.10.: Mean and standard deviation of residuals from overall latent variable model in training sets

	One Latent Variable		Two Latent Variables		Three Latent Variables	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Heidelberg Men	0.014	1.429	-0.003	1.269	0.002	1.194
Heidelberg Women	0.010	1.260	0.018	1.153	0.019	1.125
Malmo Men	-0.005	1.574	0.013	1.445	0.007	1.382
Malmo Women	0.025	1.494	0.004	1.383	0.007	1.306
Graz Men	0.004	1.276	-0.007	1.183	-0.006	1.131
Graz Women	-0.025	1.099	-0.025	1.099	-0.021	1.030

Table 7.11.: Mean and standard deviation of residuals from latent variable models in testing populations using own training set model

dard deviations are even slightly smaller in many cases using this model compared to the centre-specific model: see table 7.12. We can therefore conclude that the overall model should be as good as the centre-specific model for identifying deformities.

7.3.7. Identification of Deformed Vertebrae

All models are extremely good at identifying deformed vertebrae. However, the best discrimination was offered by the single factor model: adding further factors actually reduced the area under the ROC curve (see Table 7.13).

7.3.8. Identification of Subjects with Deformities

Again, both methods are good at identifying subjects with deformed vertebrae. Using the relative, rather than absolute, reduction in height appears to work better,

	One Latent Variable		Two Latent Variables		Three Latent Variables	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
Heidelberg Men	-0.013	1.417	0.002	1.243	0.003	1.184
Heidelberg Women	0.036	1.265	0.022	1.122	0.023	1.071
Malmo Men	-0.017	1.572	0.010	1.416	0.009	1.335
Malmo Women	0.052	1.533	0.013	1.351	0.012	1.281
Graz Men	-0.030	1.351	-0.007	1.216	-0.007	1.164
Graz Women	-0.011	1.220	-0.011	1.084	-0.010	1.008

Table 7.12.: Mean and standard deviation of residuals from latent variable models in testing sets using combined model

Variable	One Factor	Two Factors	Three Factors
Min	0.9733	0.9733	0.9659
RelMin	0.9842	0.9838	0.9763

Table 7.13.: Area under ROC's for prediction of deformed vertebrae

and there is a slight improvement by using three factors: see table 7.14.

Variable	One Factor	Two Factors	Three Factors
Min	0.9248	0.9177	0.9219
RelMin	0.9558	0.9574	0.9623

Table 7.14.: Area under ROC's for prediction of subjects with deformed vertebrae

7.4. Discussion

Using a single latent variable is similar to assuming all spines are the same shape and differ only in size[1]. Thus it gives results very similar to the polynomial model. However, modelling the natural variation in shape of the spine by including more than one latent variable leads to a significant improvement in prediction.

Considering each population separately may not be the best way to define a latent variable model. If a particular latent variable varies little within a population, it will not be detected as important in the factor analysis of that population. However, it may well vary considerably either within a different population or between populations, in which case it will be important to include it in any model of vertebral heights.

It is common practice in factor analysis to rotate the factors in order to obtain vectors that are easier to interpret. However, in this instance, the rotated vectors

[1]It does not correspond exactly: it would if the first latent factor was equal to the vector of mean heights

were harder to interpret than the unrotated vectors presented. In addition, the predicted heights from the factor analysis model would not be affected by the rotation, so the results of the rotation are not given.

Unfortunately, latent variable models do not work well if there are missing values. Since all measured heights are used to calculate the factor scores, if any height is missing the factor scores cannot be calculated. Hence no prediction can be made. In theory, it is possible to replace the missing values by predicted values from one of the other methods, or even start with such predicted values and apply the latent variable model iteratively to end up with predicted values for all heights from that model.

However, there is a more serious problem in that the latent variable models are not robust. The factor scores are calculated based on all of the measured heights, and we know that if there is a fracture in the spine, some of the heights will be reduced in that vertebra. The lack of robustness in the latent variable analysis is a serious problem. It is possible to use a robust estimate of the correlation matrix to ensure that the factors themselves are unbiased, but it is not possible to produce unbiassed estimates of the heights. For this reason, the latent variable models will not be pursued further.

8. Polynomial Models of the Spine with Varying Magnification

The models in Chapters 5 and 6 assumed that all spines within a given population were the same shape, and varied only in size. However, we saw in Chapter 7 that this is not the case: latent variable models that allowed for differences in shape between spines provided a markedly better fit to the unfractured spines than models that did not. We would like a way to incorporate such differences in shape into a polynomial model, since we have seen that the polynomial models have advantages in terms of robustness and coping with missing data.

8.1. Model Fitting

The basic polynomial model with a single magnification factor per subject is

$$H_{sij} = m_j \times \left(\sum_{k=0}^{k=r} a_{sk} \times i^k \right) + \epsilon_{sij} \tag{8.1}$$

In this model, the parameters a_{sk} define the shape of the spine, and m_j its size. We have seen that although the vertebral heights always increase in size from T4 to L4, the extent to which they increase differs between individuals. Therefore, it may be that allowing only a single magnification factor (m_j) for each subject is not sufficient to capture the natural variation in shape between individuals.

8.1.1. Linearly Increasing Magnification

The simplest way to allow more natural variation in shape is to fit the model

$$H_{sij} = (m_{1j} + m_{2j} \times (i-1)) \times \left(\sum_{k=0}^{k=r} a_{sk} \times i^k \right) + \epsilon_{sij} \tag{8.2}$$

In this model, m_{1j} measured the magnification at T4, whilst m_{2j} measures the increase in magnification for each succeeding vertebra. The main reason for choosing this model is its simplicity, but there is some justification for it in the latent variable models. In these models, the second most important difference between spines was the extent to which the vertebrae increased in height between T4 and L4. Although assuming a linear increase in magnification is likely to be an oversimplification, we would expect some improvement in fit from this model.

In order to fit this model, two new variables were calculated:

$$x_1 = \sum_{k=0}^{k=r} a_{sk} \times i^k$$

and

$$x_2 = (i-1) \left(\sum_{k=0}^{k=r} a_{sk} \times i^k \right)$$

Then robust regression was used to fit the model

$$H_{sij} = m_{1j}x_1 + m_{2j}x_2 + \epsilon_{sij} \tag{8.3}$$

which is equivalent to Equation 8.2. This model will be referred to as the linear magnification increase model, since the magnification increases linearly along the spine.

8.1.2. Two Distinct Magnification Factors

An alternative to this linear increase in magnification along the spine is to fit two distinct magnification factors: one to the lower spine and one to the upper spine. This mimics the manner in which the data was collected: two films were required to visualise all 13 vertebrae, and it is possible that the two films were at slightly different magnifications. The point at which the change in magnification occured is unknown (several vertebrae appeared on both films and were read from whichever film had the clearer image). We are therefore fitting the model

$$H_{sij} = \begin{cases} m_{1j} \times \left(\sum_{k=0}^{k=r} a_{sk} \times i^k\right) + \epsilon_{sij}, & i \leq v_j \\ (m_{1j} + m_{2j}) \times \left(\sum_{k=0}^{k=r} a_{sk} \times i^k\right) + \epsilon_{sij}, & i > v_j \end{cases} \tag{8.4}$$

in which m_{1j} is the magnification of the upper v_j vertebrae and m_{2j} the change in magnification of the lower $13 - v_j$. Thus there are three parameters per individual in this model (m_{1j}, m_{2j} and v_j), compared to two parameters per person in the linearly increasing magnification model.

The two new variables that needed to be calculated to fit this model were

$$x_1 = \sum_{k=0}^{k=r} a_{sk} \times i^k$$

and

$$x_2 = \begin{cases} 0, & i \leq v_j \\ \left(\sum_{k=0}^{k=r} a_{sk} \times i^k\right), & i > v_j) \end{cases}$$

However, v_j is not known, and needs to be estimated for each subject. Therefore, x_2 was calculated for values of v_j from 3 to 10, since the upper three vertebrae did not appear on the lower film and the lower three vertebrae did not appear on the upper film. Having at least three vertebrae at each magnification reduced the probability that a fracture at T4 or L4 would be missed because a separate magnification factor was fitted to that vertebra. The root mean square error (RMSE) for each individual was calculated for each value of v_j, and the value of v_j with the lowest RMSE was chosen for that individual. This model will be referred to as the categorical magnification increase model, since the vertebrae in a given subject can be divided into two groups with different magnifications.

8.2. Results

8.2.1. Comparison to Robust Polynomial Model in Training Sets

The mean and standard deviation of the residuals from both linear and categorical magnification models are given in Table 8.1, along with the results from the single magnification factor model.

It can be seen that the standard deviation of the residuals is smaller for the varying magnification models. There is a slight bias in these models, with the heights being underestimated on average, but the difference is never more than

	Single Factor		Categorical		Linear	
Population	Mean	SD	Mean	SD	Mean	SD
Heidelberg Men	-0.00	1.49	-0.05	1.38	-0.03	1.36
Heidelberg Women	-0.00	1.24	-0.02	1.15	-0.01	1.13
Malmo Men	0.00	1.50	-0.03	1.40	-0.03	1.40
Malmo Women	-0.00	1.36	-0.02	1.25	-0.02	1.24
Graz Men	0.00	1.20	-0.01	1.13	-0.01	1.12
Graz Women	0.00	1.22	-0.02	1.13	-0.02	1.13

Table 8.1.: Means and standard deviations of residuals from fixed and varying magnification polynomial models in training sets

0.05mm. Thus we can conclude that the varying magnification models fit these populations better than the fixed magnification model.

8.2.2. Comparison to Robust Polynomial Model in Testing Sets

Table 8.2 gives the same results as Table 8.1, but for the testing samples rather than the training samples.

Again, the varying magnification models have smaller standard deviations than the single magnification factor model. There are again slight biases, this time for all three models, but again less than 0.05mm in all cases.

	Single Factor		Categorical		Linear	
Population	Mean	SD	Mean	SD	Mean	SD
Heidelberg Men	-0.01	1.45	-0.03	1.31	-0.02	1.30
Heidelberg Women	0.01	1.23	-0.03	1.13	-0.02	1.13
Malmo Men	0.01	1.58	-0.02	1.46	-0.03	1.46
Malmo Women	-0.03	1.46	-0.03	1.36	-0.02	1.36
Graz Men	-0.00	1.32	-0.04	1.20	-0.03	1.20
Graz Women	0.02	1.21	-0.01	1.09	-0.01	1.09

Table 8.2.: Means and standard deviations of residuals from fixed and varying magnification polynomial models in testing sets

8.2.3. Performance of Varying Magnification Models in Identifying Deformities

The thresholds required to give the same sensitivity and specificity as the McCloskey-Kanis model using the linearly and categorically varying magnification models are given in Table 8.3. The thresholds to give the same sensitivity are slightly less than those to give the same specificity. This suggests that the models will be slightly less good, since the less stringent threshold suggests that the specificity will be poorer.

	Sensitivity	Specificity
Categorical	15.5%	16.2%
Linear	15.4%	16.1%

Table 8.3.: Thresholds for varying magnification models

The numbers of vertebrae classed as deformities by these two models are given in Tables 8.4 and 8.5.

8.2.4. Performance of Varying Magnification Models in Identifying Subjects with Deformities

Table 8.6 shows the classification of subjects as cases or non-cases using both varying magnification models. Several thresholds were used for each model: 15%, 20%, and two additional thresholds, one to give the same sensitivity as the McCloskey-Kanis method and one to give the same specificity as the McCloskey-Kanis method.

8.3. Discussion

There are good *a priori* reasons for using either of the varying magnification models outlined in this chapter. The linearly increasing magnification was suggested by the latent variable models, in which the second most important source of variation between individuals was the extent to which the lower vertebrae were larger than the upper vertebrae.

On the other hand, we knew that two films were required to measure all 13 vertebrae, but not which vertebrae were measured on which film. Fitting two distinct magnification factors and allowing the vertebra at which the magnification factor changed to be determined by the data modelled this process.

There was little difference in how well these two models fitted the data from subjects without deformities. Both models fitted better than any of the models we

Population	Model Type	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	Categorical	9/910=1.0%	3/701=0.4%	6/271=2.2%	20/35=57.1%	9/476 =1.9%
Heidelberg Men	Linear	13/910=1.4%	3/701=0.4%	4/271=1.5%	20/35=57.1%	10/476=2.1%
Heidelberg Women	Categorical	2/910=0.2%	1/811=0.1%	4/208=1.9%	42/49=85.7%	4/139=2.9%
Heidelberg Women	Linear	3/910=0.3%	3/811=0.4%	5/208=2.4%	44/49=89.8%	4/139=2.9%
Malmo Men	Categorical	7/910=0.8%	12/1022=1.2%	7/439=1.6%	53/61=86.9%	22/209=10.5%
Malmo Men	Linear	7/910=0.8%	8/1022=0.8%	6/439=1.4%	49/61=80.3%	20/209=9.6%
Malmo Women	Categorical	6/910=0.7%	10/1218=0.8%	9/450=2.0%	55/62=88.7%	7/114=6.1%
Malmo Women	Linear	7/910=0.8%	10/1218=0.8%	9/450=2.0%	55/62=88.7%	6/114=5.3%
Graz Men	Categorical	3/910=0.3%	6/1025=0.6%	3/389=0.8%	41/54=75.9%	5/241=2.1%
Graz Men	Linear	2/910=0.2%	5/1025=0.5%	1/389=0.3%	47/54=87.0%	7/241=2.9%
Graz Women	Categorical	7/910=0.8%	4/1391=0.3%	1/225=0.4%	19/30=63.3%	1/104=1.0%
Graz Women	Linear	2/910=0.2%	5/1391=0.4%	1/225=0.4%	21/30=70.0%	1/104=1.0%
Total	Categorical	34/5460 = 0.6%	36/6168 = 0.6%	30/1982=1.5%	230/291=79.0%	48/1283=3.7%
Total	Linear	34/5460 = 0.6%	34/6168 = 0.6%	26/1982=1.3%	236/291=81.8%	48/1283=3.7%

Table 8.4.: Number of deformities from varying magnification models with the same specificity as McCloskey-Kanis model in training samples

Population	Model Type	Training Set (All Verts)	Other Normals (All Verts)	Fractures (Norm. Verts)	Fractures (Fx Verts)	Other Deformities (All Verts)
Heidelberg Men	Categorical	11/910=1.2%	4/701=0.6%	6/271=2.2%	21/35=60.0%	11/476 =2.3%
Heidelberg Men	Linear	13/910=1.4%	3/701=0.4%	5/271=1.9%	20/35=57.1%	11/476=2.3%
Heidelberg Women	Categorical	2/910=0.2%	2/811=0.3%	4/208=1.9%	44/49=89.8%	4/139=2.9%
Heidelberg Women	Linear	3/910=0.3%	3/811=0.4%	7/208=3.4%	44/49=89.8%	5/139=3.6%
Malmo Men	Categorical	9/910=1.0%	14/1022=1.4%	8/439=1.8%	55/61=90.2%	22/209=10.5%
Malmo Men	Linear	7/910=0.8%	9/1022=0.9%	7/439=1.7%	51/61=83.6%	20/209=9.6%
Malmo Women	Categorical	6/910=0.7%	12/1218=1.0%	10/450=2.2%	56/62=90.3%	7/114=6.1%
Malmo Women	Linear	7/910=0.8%	10/1218=0.8%	10/450=2.2%	56/62=90.3%	7/114=6.1%
Graz Men	Categorical	4/910=0.4%	7/1025=0.7%	3/389=0.8%	44/54=81.5%	5/241=2.1%
Graz Men	Linear	3/910=0.3%	7/1025=0.7%	2/389=0.5%	48/54=88.9%	7/241=2.9%
Graz Women	Categorical	10/910=1.1%	5/1391=0.4%	1/225=0.4%	20/30=66.7%	2/104=1.9%
Graz Women	Linear	4/910=0.4%	5/1391=0.4%	1/225=0.4%	21/30=70.0%	2/104=1.9%
Total	Categorical	42/5460 = 0.8%	44/6168 = 0.7%	32/1982=1.6%	240/291=82.5%	51/1283=4.0%
Total	Linear	37/5460 = 0.7%	37/6168 = 0.6%	32/1982=1.6%	240/291=82.5%	52/1283=4.1%

Table 8.5.: Number of deformities from varying magnification models with the same sensitivity as McCloskey-Kanis model in training samples

Method	Threshold	Training Set	Testing Set	Fractures	Other Deformities
Linear	15.0%	24/420	27/489	136/178	36/105
	15.4%	19/420	23/489	133/178	35/105
	16.1%	18/420	20/489	127/178	33/105
	20.0%	6/420	5/489	95/178	16/105
Categorical	15.0%	26/420	32/489	135/178	38/105
	15.5%	24/420	27/489	135/178	34/105
	16.2%	21/420	23/489	125/178	31/105
	20.0%	6/420	2/489	92/178	19/105

Table 8.6.: Cross-tabulation of morphometric and clinical classifications of the presence of at least one deformity in a subject

have seen so far.

The detection of deformities was also good, being virtually identical in performance to the McCloskey-Kanis method. However, the McCloskey-Kanis method was better at identifying subjects with deformities.

9. Outlier Detection Methods

9.1. Introduction

We have seen, particularly in chapter 7, that there is a considerable amount of structure in the vertebral height measurements. The data from subjects without deformities lie in a very small subset of the 39-dimensional space of all possible heights. Fractures reduce the height of one or more vertebrae considerably, but leave other heights unchanged, thus changing the overall shape of the spine. It is therefore reasonable to assume that such points may appear to be outliers from the bulk of the data.

If we knew that the vertebral heights followed a known distribution, it would be relatively straightforward to identify outlying observations, based on the known mean vector and covariance matrix. We could calculate the mean vector and covariance matrix based on the subjects known to be fracture free, and use these values to determine outliers. However, there are a number of drawbacks to this method:

1. Different populations may need different mean vectors and covariance matrices, but the numbers of subjects available to calculate them from are quite small.

2. This method identifies subjects with fractures, rather than fractured vertebrae. Therefore, the fact that we need to know who has fractures and who does not before we can apply the method makes it unusable.

The alternative is to use robust methods to calculate the mean vector and covariance matrix from the entire dataset. This is a difficult problem if there are multiple outliers, since they can affect the mean vector and covariance matrix sufficiently that they do not appear as outliers. However, one approach to identifying multivariate outliers developed by Hadi [23, 24] and implemented in Stata was investigated.

9.2. Multivariate Outlier Detection

According to Rousseeuw and van Someren [25], outliers are 'observations that do not follow the pattern of the majority of the data'. We can follow Rocke and Woodruff [17] in thinking of the data being divided up into 'good' data and 'bad' data. The 'good' data is that which comes from the parent distribution that we are interested in (in our case height measurements from the vertebrae of normal spines), and the 'bad' data is any data in our sample from a different distribution (height measurements from the vertebrae of abnormal spines).

Therefore, we need to

1. Define what is normal. This involves estimating the location and shape of the distribution of 'good' data.

2. Use these estimates to generate a suitable statistic to distinguish between 'good' and 'bad' data, and class those points sufficiently far from the location of the 'good' data as 'bad' data.

Suppose that normal vertebral heights follow a p-dimensional multivariate normal distribution with mean $\mathbf{M}$ and covariance matrix $\mathbf{\Sigma}$. If we know $\mathbf{M}$ and $\mathbf{\Sigma}$, we test whether a given observation $\mathbf{x_i}$ was an outlier using the Mahalonobis distance

$$MD_i = (\mathbf{x_i} - \mathbf{M})'\mathbf{\Sigma}^{-1}(\mathbf{x_i} - \mathbf{M}) \tag{9.1}$$

The MD_i would follow a χ^2 distribution on p degrees of freedom, from which a suitable threshold c could be chosen. Any observation with $MD_i > c$ would be classed as an outlier. The specificity of the method can be determined by the choice of c: if $c = \chi^2_{p;1-\alpha}$, then the specificity is $1 - \alpha$. In other words, a proportion α of normal observations will be classed as outliers using this choice of threshold (Note that some authors [26] use $c = \chi^2_{p;1-\alpha/n}$ as the threshold, where n is the sample size, since this fixes the error-rate *per experiment* at α. We prefer to fix the error-rate *per subject*, since this reflects the commonly accepted definition of specificity. This approach also seems more common in recent papers, such as [27]. Hadi appears to recommend $1-\alpha$[23], but the implementation of his algorithm in Stata used $1-\alpha/n$. I modified the stata code to use $1 - \alpha$).

Unfortunately, in general, $\mathbf{\Sigma}$ and $\mathbf{M}$ are not known, and need to be estimated from the data. The conventional estimators for $\mathbf{M}$ and $\mathbf{\Sigma}$, are

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p) \tag{9.2}$$

where

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{9.3}$$

and

$$\mathbf{S} = \frac{(\mathbf{X} - \mathbf{M})'(\mathbf{X} - \mathbf{M})}{\mathbf{n} - \mathbf{1}} \tag{9.4}$$

However, these estimators are not robust, and can be greatly affected by outliers. Using these estimators to calculate MD_i can lead to errors in identifying outliers known as *masking* and *swamping*

Masking is caused when there are a cluster of outliers. They attract $\bar{\mathbf{x}}$ and inflate $\mathbf{S}$ in their direction. This will lead to MD_i being underestimated in this group, and it may be that none of the cluster of outliers is detected. The term masking refers to the presence of one outlier masking the appearance of another.

Since the cluster of outliers has attracted $\bar{\mathbf{x}}$ and inflated $\mathbf{S}$ in their direction, MD_i for observations lying in the opposite direction will be overestimated. In serious cases, the effect may be large enough for normal observations to be declared as outliers. This effect is known as swamping.

Hence, to use the Mahalanobis distance to identify outliers, robust estimators for $\mathbf{\Sigma}$ and $\mathbf{M}$ are required. Hadi refers to distances analogous to the Mahalanobis distance but calculated using robust estimators for $\mathbf{\Sigma}$ and $\mathbf{M}$ as *robust distances*, abbreviated to RD_i. There are a wide variety of possible robust estimators: combinatorial estimators, which rely on identifying a small subset of observations that do not contain outliers; M- and S-estimators, which weight observations according to

how 'unusual' they are; and sequential point addition estimators, which begin with a small 'normal' subset and then add the continue to add the 'next most normal' point until all points have been added or there is no point left that is sufficiently 'normal' to be added.

The method we chose to use was a sequential point addition estimator, due to Hadi [23]. This method begins by calculating RD_i using very robust (but inefficient) estimators for M and Σ. Then the observations are sorted into increasing order of RD_i, and the first $r = \frac{n+p+1}{2}$ are referred to as the 'basic subset' (this value of r is that used in the stata implementation: in Hadi's original paper, a value of $p+1$ was used). The new robust estimator for M is the sample mean vector of this basic subset, m_b, and the new robust estimator for Σ is given by $c_{np}S_b$, where S_b is the sample covariance matrix for the basic subset and $c_{np} = (1 + \frac{2}{n-1-3p} + \frac{p+1}{n-p})^2$ is an empirical small sample correction factor. These estimators are now substituted into equation 9.1 to recalculate the RD_i. The value of RD_{r+1} (i.e. the distance of the smallest observation that was not included in the basic subset) is then compared to $\chi^2_{p;1-\frac{\alpha}{n}}$, where α is a significance level chosen by the user. If RD_{r+1} is less than this threshold, the observation is added to the basic subset, and new estimators and RD_i are calculated. If RD_{r+1} is greater than this threshold, all observations not currently in the basic subset are declared to be outliers. Thus S_b and m_b are again calculated as the sample mean and covariance matrix of a subset of the n observations (up to a correction factor for S_b). This method has been implemented in stata as the procedure `hadimvo`.

9.3. Methods

9.3.1. Missing Data

There are two possible approaches to subjects with missing data. One is to ignore the variables for which the subject has no data. This can be done by dropping the corresponding rows and columns of the mean vector and covariance matrix. Thus, the problem is reduced from 39 dimensions to $39 - m$ dimensions, where m is the number of missing values. The calculated distance RD would then be compared to a χ^2 distribution on $39 - m$ degrees of freedom to identify outliers.

Alternatively, missing values could be replaced by predicted values, using one of the prediction methods outlined in the previous chapters. However, since the natural variation in vertebral heights is not included in the predicted heights, the distance RD would no longer follow a χ^2 distribution on 39 degrees of freedom. Therefore, the first method was of handling missing values was used.

Subjects with missing values were not used to determine $\bar{\mathbf{x}}$ and $\mathbf{S}$. However, when $\bar{\mathbf{x}}$ and $\mathbf{S}$ have been calculated, the appropriate rows and columns can be used to determine RD from the observed measurements. RD can then be compared to a χ^2-distribution on r degrees of freedom, where r is the number of non-missing measurements, to provide a p-value. This p-value is used to determine which observations are classed as outliers and which are not. In order to obtain a value for RD which is comparable to the values of RD calculated for the the other subjects, this p-value can then be converted into a distance by comparing it to a χ^2-distribution on 39 degrees of freedom. In this way, RD from the subjects with missing mea-

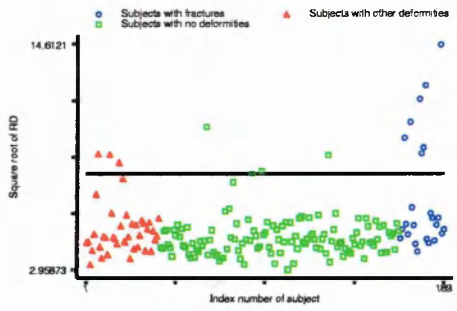
surements should be directly comparable with RD from subjects with no missing measurements.

To test whether this procedure was successful, subjects in whom all heights were known had some of their heights set to missing. Between 1 and 10 heights in each spine were set to missing at random locations in the spine. RD was then calculated using the method described above, and compared to the value of RD obtained when all the heights were known.

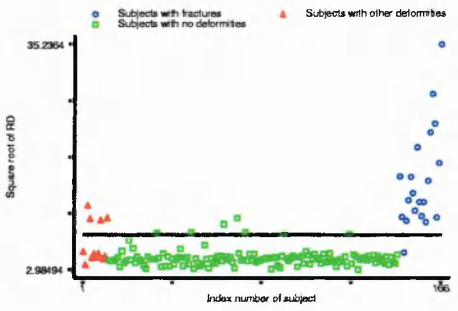
9.4. Results

The results of applying Hadi's method to the data is shown in Figures 9.1 and 9.2. It should be noted that in these graphs, the Y-axis shows the square root of RD, which is the measure returned by the procedure `hadimvo` in stata. Two values of α were used: 0.01 and 0.05. The value of 0.05 was chosen since approximately 5% of normal subjects were classed as having at least one prevalent fracture using the McCloskey-Kanis method. If the measured heights follow a multivariate normal distribution, the specificity of this outlier method with $\alpha = 0.05$ should be similar to the specificity of the McCloskey-Kanis model. Figures 9.1 and 9.2 both appear to have achieved a clear separation between the observation classed as 'normal' and and those classed as outliers outliers, despite the fact that more observations were classed as outliers using $\alpha = 0.05$ than using $\alpha = 0.01$. This apparent clear separation between normal observations and outliers is in fact artefactual, and explained in the section 9.5.1.

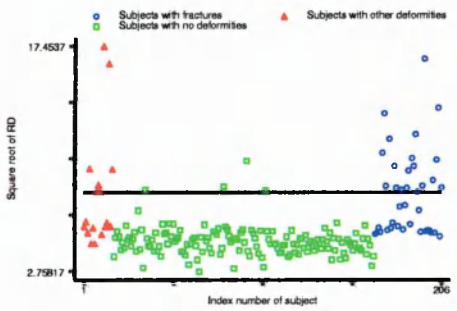
Tabulations of the number of subjects classed as having at least one deformity



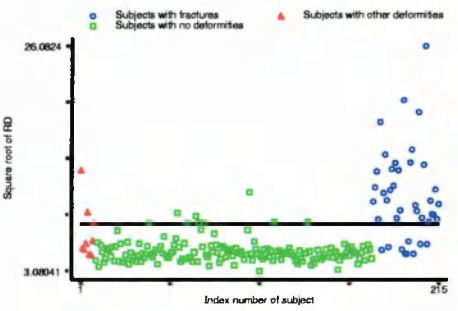
(a) Men in Heidelberg



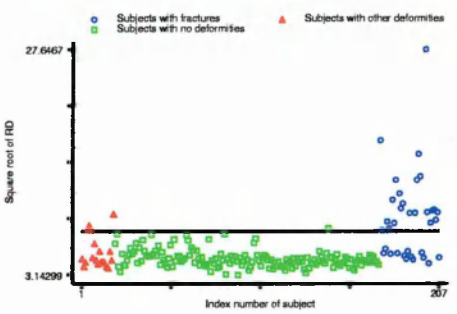
(b) Women in Heidelberg



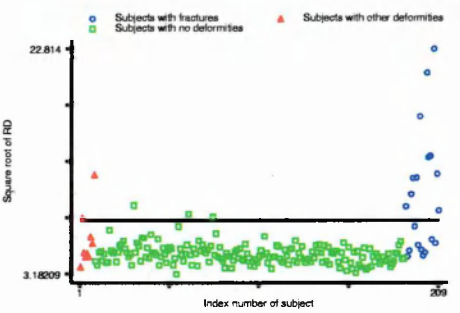
(c) Men in Malmo



(d) Women in Malmo

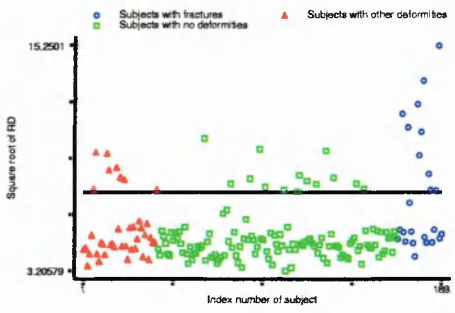


(e) Men in Graz

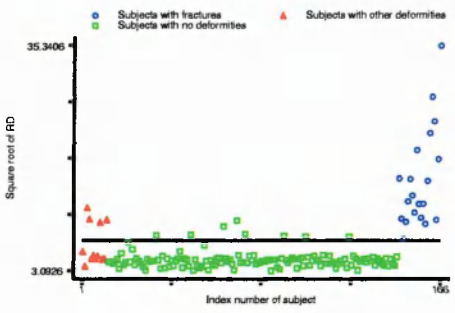


(f) Women in Graz

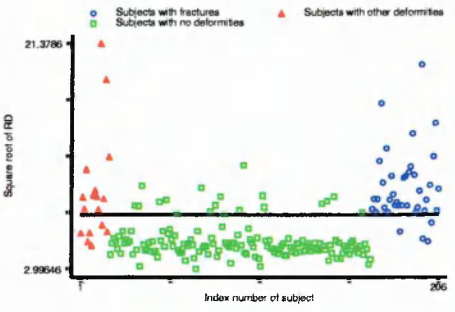
Figure 9.1.: Outliers using Hadi's methods with $\alpha = 0.01$



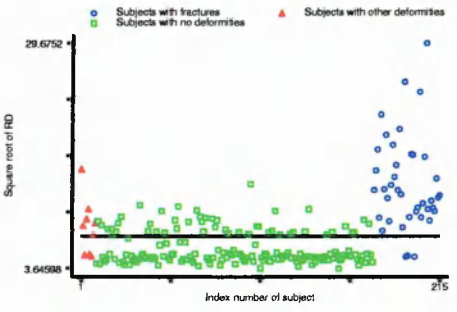
(a) Men in Heidelberg



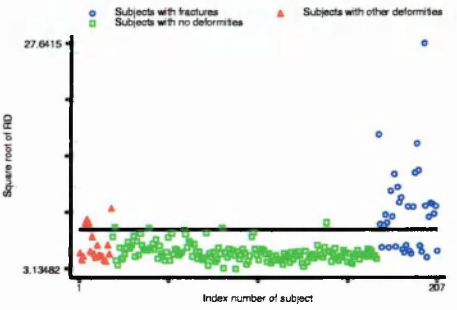
(b) Women in Heidelberg



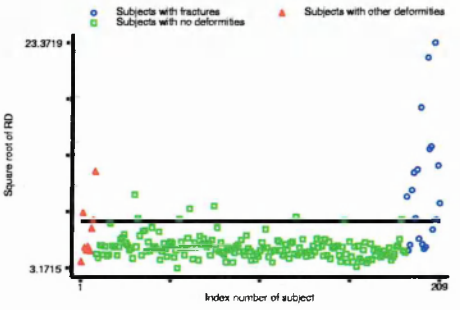
(c) Men in Malmo



(d) Women in Malmo



(e) Men in Graz



(f) Women in Graz

Figure 9.2.: Outliers using Hadi's methods with $\alpha = 0.05$

in each of the subpopulations are given in Tables 9.1 and 9.2.

Population	Training Set	Other Normals	Subjects with	
			Fractures	Other Deformities
Heidelberg Men	1/70=1.4%	0/43=0%	7/21=33.3%	0/28=0%
Heidelberg Women	2/70=2.9%	3/54=5.6%	17/18=94.4%	2/9=22.2%
Malmo Men	1/70=1.4%	1/65=1.5%	17/33=51.5%	4/11=36.4%
Malmo Women	2/70=2.9%	4/79=5.1%	27/36=75.0%	2/7=28.6%
Graz Men	1/70=1.4%	0/56=0%	19/31=61.3%	2/16=12.5%
Graz Women	0/70=0%	1/92=1.1%	10/17=58.8%	0/6=0%
All Men	3/210=1.4%	1/164=0.6%	43/85=50.6%	6/55=10.9%
All Women	4/210=1.9%	8/226=3.5%	54/71=76.1%	4/22=18.2%

Table 9.1.: Numbers of subjects classified as outliers in each group using Hadi's method with $\alpha = 0.01$

It is clear from the above tables that the proportion of deformity-free subjects classed as outliers by this method is greater than the nominal proportion in most populations. The tables also show that using the same value of α for both men and women may not be a good idea, since for a given value of α, the sensitivity is greater and the specificity lower in women than in men.

Again, the difference in sensitivity suggests that the method may be better at detecting some types of fracture than others, since a greater proportion of fractures

Population	Training Set	Other Normals	Subjects with	
			Fractures	Other Deformities
Heidelberg Men	10/70=14.3%	0/43=0%	10/21=47.6%	3/28=10.7%
Heidelberg Women	3/70=4.3%	3/54=5.6%	18/18=94.4%	2/9=22.2%
Malmo Men	6/70=8.6%	10/65=15.4%	30/33=90.9%	8/11=72.7%
Malmo Women	17/70=24.3%	22/79=27.9%	33/36=91.7%	4/7=57.1%
Graz Men	1/70=1.4%	0/56=0%	20/31=64.5%	2/16=12.5%
Graz Women	3/70=4.3%	2/93=1.1%	11/17=64.7%	1/6=16.7%
All Men	17/210=8.1%	10/164=6.1%	60/85=70.6%	13/55=23.6%
All Women	23/210=10.9%	27/226=11.9%	62/71=87.3%	7/22=31.8%

Table 9.2.: Numbers of subjects classified as outliers in each group using Hadi's method with $\alpha = 0.05$

in men are the less severe wedge fractures. Therefore each subject was classified according to the most severe type of fracture they had (none, wedge, concavity, biconcavity or crush). The proportion of subjects in each of these groups classified as an outlier are given in Table 9.3.

	Worst Fracture	Men	Women	Combined
$\alpha = 0.01$	None	$10/429 = 2.3\%$	$16/458 = 3.5\%$	$26/887 = 2.9\%$
	Wedge	$19/52 = 36.5\%$	$11/26 = 42.3\%$	$30/78 = 38.5\%$
	Concavity	$20/27 = 74.1\%$	$34/36 = 94.4\%$	$54/63 = 85.7\%$
	Biconcavity	$3/3 = 100\%$	$6/6 = 100\%$	$9/9 = 100\%$
	Crush	$1/1 = 100\%$	$2/2 = 100\%$	$3/3 = 100\%$
$\alpha = 0.05$	None	$40/429 = 9.3\%$	$57/458 = 12.5\%$	$97/887 = 10.9\%$
	Wedge	$30/52 = 57.7\%$	$18/26 = 69.2\%$	$48/78 = 61.5\%$
	Concavity	$25/27 = 92.6\%$	$35/36 = 97.2\%$	$60/63 = 95.2\%$
	Biconcavity	$3/3 = 100\%$	$6/6 = 100\%$	$9/9 = 100\%$
	Crush	$1/1 = 100\%$	$2/2 = 100\%$	$3/3 = 100\%$

Table 9.3.: Percentage of subjects classified as outliers according to severity of their worst deformity

Table 9.3 shows that as with the other methods, the proportion of fractures correctly identified (the sensitivity of the method) increases with the severity of the fracture. The sensitivity to any given type of fracture did not differ significantly between men and women, and hence the difference in sensitivity between men and women was due to the different proportions of the different types of fractures.

9.4.1. Merging Populations

It would be an advantage if we could merge the different populations, and simply search for outliers in a single population, rather than in six subpopulations. However, the fact the different sub-populations showed considerable differences in various respects in Chapter 7 suggests that this may not work. To test the possibility, the method was applied to men and women separately. The proportion of subjects classed as outliers in each of the subgroups is given in Table 9.4.

Gender	α	Training Set	Other Normals	Fractures	Other Deformities
Men	0.01	24/210=11.4%	19/164=11.6%	62/85=72.9%	15/55=27.3%
	0.05	68/210=32.4%	55/164=33.5%	77/85=90.6%	30/55=54.5%
Women	0.01	18/210= 8.5%	22/226=9.7%	60/71=84.5%	6/22=27.3%
	0.05	75/210=35.7%	67/226=27.7%	66/71=92.7%	13/22=59.1%

Table 9.4.: Number of subjects classed as outliers when populations are merged.

Table 9.4 clearly shows that merging the populations has been unsuccessful. The proportions of false positives among both men and women is extremely high for both choices of α. This could have been anticipated, since we saw in Chapter 7 that factor analysis gave very different results in each centre, suggesting that both M and Σ differ between centres. Using a single estimate of M and Σ for all centres will therefore be a poor way to identify outliers.

9.4.2. Dealing with Missing Values

There was a very strong correlation between RD, calculated with all heights measured, and RD_m, calculated with some heights missing ($r = 0.95$). This is shown in Figure 9.3. However, RD_m tended to be greater than RD, and to increase with an increasing number of missing values.

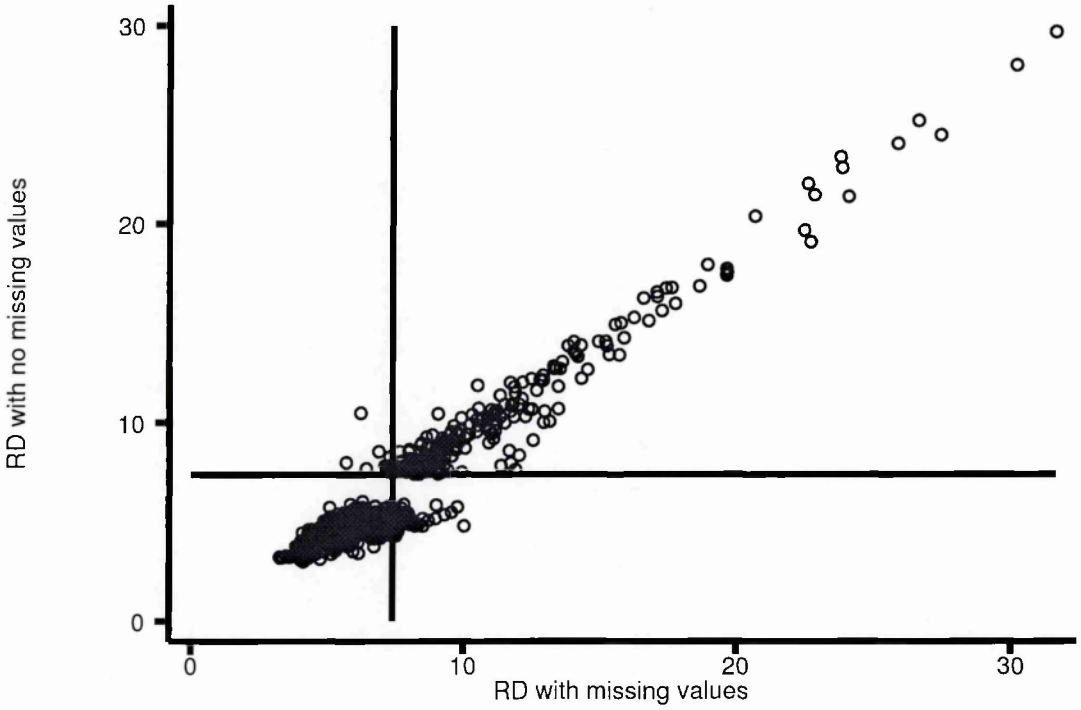


Figure 9.3.: Correlation between RD with and without missing values

This suggests that there will be more outliers using RD_m than using RD, and this is in fact the case. Of the 1043 subjects with complete data, 219 (21%) were classed as outliers when the complete data were used, and 254 (24%) when the missing data was used. This increase was statistically significant using McNemar's test. In subjects classed as normal by the radiologist, the false positive rate increased from

9.5% to 13.2%, $p < 0.0001$. However, in subjects with fractures, the true positive rate only increased from 78.2% to 78.8%, which was not significant. The increase in the 77 subjects with deformities other than fractures was from 26% to 31%, which was again not significant.

9.5. Discussion

9.5.1. Hadi's Outlier Detection Algorithm

One unusual aspect of this approach is that the observations are divided up into one sub-sample containing 'normal' observations, and the other containing 'outliers'. One might expect the sample covariance matrix of the normal observations to be a reasonable estimate of the covariance matrix in the 'normal' population, but this is not the case: a multiplying factor of $c_{np} = (1 + \frac{2}{n-1-3p} + \frac{p+1}{n-p})^2$ is required.

In fact, using this factor does not give a good estimate of the population covariance matrix, but it does discriminate well between 'normal' observations and outliers. This is because RD_i does not follow a χ^2 distribution in small samples, since S and $\bar{x}$ are not independent of x_i. It has been shown [26] that in the basic subset,

$$RD_i \sim \frac{p(n-1)^2 F_{p,n-p-1}}{n(n-p-1+pF_{p,n-p-1})} \tag{9.5}$$

However, in the 'non-basic' subset, if x_i is drawn from the normal population, Krzanowski [28] has shown that

$$RD_i \sim \frac{p(n^2-1)}{n(n-p)}) F_{p,n-p} \tag{9.6}$$

where n is the size of the basic subset in both equations above.

Empirically, the distribution of RD_i in equation 9.6 is very similar to $c_{np}\chi^2_p$ and hence the number of 'normal' observations classed as outliers is appropriate to the threshold chosen. However, if the method were used in a multivariate normal population without outliers, the expected value of S would be Σ, the population covariance matrix. Thus, S is a better estimate of Σ than $c_{np}S$.

More worrying is the fact that the smallest value that may be expected in the non-basic subset is greater than the largest value expected in the basic subset. The difference can be quite marked in small samples. Hence the suggestion in [23] that outliers may be identified from consideration of an index plot of the RD_i is unfortunate: there will be a clear band on the plot between those observations included in the basic subset and those excluded, which will tend to exaggerate the difference between the two subsets. It may be that a slight change in one of the outliers would result in it no longer being classed as an outlier, whilst the index plot makes it appear that a considerable change would be necessary.

This is illustrated in Figure 9.4. This shows the 95^{th} centile of the distributions of RD_i in the basic and non-basic subsets, as a function of the sample size (calculated using Equations 9.5 and 9.6, as well as the 95^{th} centile of the χ^2 distribution on 39 degrees of freedom. Clearly, both distributions of RD_i converge asymptotically to the χ^2 distribution, but for modest sample sizes the difference can be marked. For sample sizes of 200 to 300 subjects, as we have, the distribution of RD_i in the basic subset is close to a χ^2 distribution, but the distribution of RD_i among the outliers bigger than that predicted by the χ^2 distribution. This may help to explain why

there are more false positives than expected for a chosen value of α: the threshold chosen based on a χ^2 distribution is less than the threshold would be if the correct distribution of RD were used.

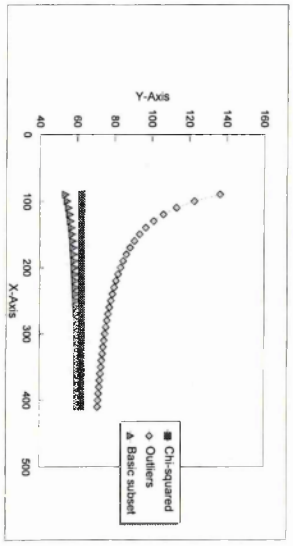


Figure 9.4.: 95^{th} centile of the distributions of RD_i in the basic and non-basic subsets

9.5.2. Missing Data

The method used for handling missing data was not very successful. There were more subjects classed as outliers when there was missing data than when the data was complete. For this reason, a more sophisticated treatment of missing values is advisable. Possibly, the imputation methods outlined in Chapter 10 could be adapted to replace missing values with imputed values. Alternatively, it may be that using a cut-off chosen from the distribution in Equation 9.6 might work better.

10. An Imputation Method Based on the McCloskey-Kanis Method

10.1. Introduction

It is a little disappointing that none of the methods produced so far have been better than the existing McCloskey-Kanis model for identifying deformities. Since the McCloskey-Kanis model only used the 4 nearest vertebrae for predicting heights, it is likely that there is greater variation in shape between spines than our models have allowed for: spines are only locally constant in shape, not globally. Indeed, this is also suggested by the improvement we have seen when allowing the magnification factor to vary within each individual.

However, there a some very obvious flaws in the McCloskey-Kanis model. We have seen that the predicted posterior height is biased, due to the method used to exclude fractured vertebrae from the calculation. Also, the predicted posterior height taken as the mean of a number of predicted values from univariate regression

models, rather than using multivariate regression, which presumably leads to a loss of precision.

Therefore, I tried to produce a model that worked in a similar way to the McCloskey-Kanis model, but using more standard statistical methodology. The great advantages of the McCloskey-Kanis model are its robustness to missing or erroneous values, so these are what needed to be reproduced. Imputation is a statistical technique widely used to replace missing data by a value predicted from the rest of the data. It can also be used to replace data believed to be unreliable (i.e. measured heights that are unusually low due to fractures) to provide a model that will provide robust estimates of vertebral heights.

10.2. Methods

The first step was to produce a predicted posterior height. This was done using the same four adjacent posterior heights as in the McCloskey-Kanis method. The intention was to use multiple regression to predict the posterior height, but this requires complete data for all subjects. Therefore any missing data needed to be replaced with imputed values.

Since there were four heights used as predictors, there were 15 possible patterns of missing data. We are only interested in 14 of them, since if all 4 heights are missing, the McCloskey-Kanis algorithm is unable to predict a posterior height, and so we do not need to either. For each possible pattern of missing values, a robust regression model (using the same weighting system as described in Chapter 6) was

used to predict the missing heights from the non-missing heights in the subjects with no missing data. Then the missing data was replaced with the predicted data from this regression model.

This provides robustness to missing data, but not to outliers. If a height is reduced due to the vertebra being fractured, all heights predicted using that height will be underestimated. To avoid this, we need to replace the measured height for a fractured vertebra with its predicted value. This will need to be done iteratively, since replacing fractured heights with their predicted values should lead to more accurate predictions for the adjacent vertebrae, and may reveal additional fractures. When no additional fractures are detected, the iteration stops. A fracture is defined as a height less than 80% of its predicted value. This procedure must be performed for each centre, sex and vertebra separately.

To determine whether the other heights in a vertebra have been affected, a similar process to the McCloskey-Kanis method was used. If the measured posterior height was less than 80% of its predicted value, only the predicted height was used to predict the anterior and mid heights, otherwise both the measured and predicted heights were used. Again, robust regression was used to determine the regression equations for each height.

A vertebra was declared a deformity if any of the three measured heights was 80% or less of its predicted value. Vertebrae were classified into shapes in the same way as the McCloskey-Kanis method, and the resulting vertebral shapes were compared to those predicted by the McCloskey-Kanis method and those assigned by the clinician.

However, the results of the McCloskey-Kanis algorithm presented in this chapter

differ from those presented earlier. This is because previously we were using a training set to determine the polynomial models, and the same training set was used to determine the reference ranges for the McCloskey-Kanis algorithm, for the sake of comparability. However, usually the entire sample is used to determine the reference ranges, so that is what has been done in this chapter.

10.3. Results

10.3.1. Accuracy of Prediction in Normal Vertebrae

The means and standard deviations of the residuals in each of the testing samples are given in Table 10.1. Comparing with Table 5.7 shows that the mean residual (bias) is markedly less with this method than with either the Minne or McCloskey-Kanis methods, but slightly higher than the polynomial model in some cases. The standard deviations, however, are lower for this model in all cases compared to the three models in Table 5.7.

10.3.2. Identification of Deformed Vertebrae

Table 10.2 gives the number of vertebrae classed as deformities using both the McCloskey-Kanis method and the new imputation method. It can be seen that both methods detect similar numbers of genuine fractures (206 for the McCloskey-Kanis method, 200 for the imputation method). However, the imputation method has considerably fewer false positives: 43, compared to 87 using the McCloskey-Kanis method. In addition, more than half of these false positives are in the subjects known

Population	Mean	S.D.
Heidelberg Men	0.08	1.30
Heidelberg Women	-0.12	1.18
Malmo Men	0.00	1.49
Malmo Women	-0.02	1.42
Graz Men	0.01	1.30
Graz Women	0.00	1.18

Table 10.1.: Mean and standard deviation of residuals in testing subgroups using imputation method.

to have deformities other than fractures in the spine. Excluding these subjects gives 20 false positives with the imputation method and 61 with the McCloskey-Kanis method.

10.3.3. Comparison of Shapes of Deformities

Table 10.3 gives a cross classification of the shape of each vertebra according to the radiologist and the shape defined by the imputation method. Clearly, the morphometric shape gives some information about the clinical shape of the deformity. Mid only deformities are most likely to be concave fractures, whilst anterior only deformities are most likely to be wedge fractures. Anterior + mid deformities are equally likely to be wedges as concavities, and post + anterior + mid deformities could be any kind of fracture, but are very rare.

A cross classification of the shape of each vertebra according to the radiologist

Vertebrae	McCloskey-Kanis	Imputation
Training Set	17/5454 = 0.3%	6/5454 = 0.1%
Testing Set	18/6141 = 0.3%	3/6143 = 0.0%
Subjects with Fracture:		
Fractured Vertebrae	206 / 287 = 71.8%	200 / 286 = 69.9%
Unfractured Vertebrae	26/1962 = 1.33%	11/1962 = 0.6%
Subjects with Other Deformities	26/1264 = 2.1%	23/1268 = 1.8%

Table 10.2.: Number of vertebrae classed as deformities using McCloskey-Kanis and imputation methods

and the shape defined by the McCloskey-Kanis method is given in Table 10.4. The results are similar to Table 10.3, although there are more false positives with most shapes of deformity. There are fewer false positives for anterior only deformities, but the proportion is similar (31% vs 33%). Interestingly, there are far more deformities in which the posterior height is affected using this method (66 vs 7), and most of them are false positive (40 out of 66, compared to 1 out of 7 using the imputation method).

10.3.4. Identification of Subjects with Deformed Vertebrae

Table 10.5 gives the number of subjects classed as having at least one vertebral deformity in each of the subject groups. The McCloskey-Kanis method is slightly more sensitive than the imputation method (135 true cases rather than 126), but far less specific (45 false positives rather than 27). Again, excluding the subjects

Morphometric Shape	Clinical Shape					
	Normal	Wedge	Concave	Biconcave	Crush	Total
Normal	14774	72	14	0	0	14860
Mid Only	5	3	44	10	0	62
Anterior Only	24	47	2	0	0	73
Anterior + Mid	11	39	39	6	6	101
Post Only	1	0	0	0	0	1
Post + Mid	0	0	0	0	0	0
Post + Anterior	0	0	0	0	0	0
Post + Anterior + Mid	0	2	1	2	1	6
Total	14819	164	100	18	7	15108

Table 10.3.: Cross-classification of vertebral shapes: imputation method

Morphometric Shape	Clinical Shape					
	Normal	Wedge	Concave	Biconcave	Crush	Total
Normal	14734	75	6	0	0	14815
Mid Only	19	11	56	10	0	96
Anterior Only	13	28	1	0	0	42
Anterior + Mid	13	33	29	5	4	89
Post Only	0	0	0	0	0	0
Post + Mid	15	0	0	0	0	15
Post + Anterior	9	0	0	0	0	9
Post + Anterior + Mid	16	12	8	3	3	42
Total	14819	164	100	18	7	15108

Table 10.4.: Cross-classification of vertebral shapes: McCloskey-Kanis method

known to have non-fracture deformities gives a greater advantage to the imputation method (9 false positives rather than 27).

Group	McCloskey-Kanis	Imputation
Training Set	12 / 420 = 2.9%	6/420 = 1.4%
Testing Set	15 / 488 = 3.1%	3/488 = 0.6%
Subjects with Fracture	135 / 177 = 76.3%	126 / 177 = 71.2%
Subjects with Other Deformities	18 / 103 = 17.5%	18/104 =17.3%

Table 10.5.: Number of subjects with deformities using the McCloskey-Kanis and imputation methods

10.4. Discussion

The McCloskey-Kanis method used in this chapter was less sensitive but more specific than that used previously. This is because previously, only the training sample was used to define the reference range, whereas in this chapter, the entire population was used. The reduced sensitivity and improved specificity was to be expected given that we know that Black's method of calculating a robust standard deviation is biased upwards in the presence of outliers.

However, the difference in reference ranges may have only been small. For the imputation method to achieve the same sensitivity as the McCloskey-Kanis method, a vertebra would have to be classed as a fracture if any of the measured heights were 16.7% less than their predicted height (rather than 20%). Using this definition, there

were 114 false positives, of which 65 were in vertebrae known to have no deformities, compared to 199 and 146 respectively with the McCloskey-Kanis method.

Changing thresholds in this way is not strictly valid, since changing the threshold would lead to different vertebrae being classed as having posterior height loss, and hence needing to be replaced by their imputed values. However, the number of vertebra affected in this way is small (8 out 15,496), and is unlikely to have an effect on the results.

A major difference between this method and the McCloskey-Kanis method is that a constant threshold of 20% was used in this model, whereas the thresholds used in the McCloskey-Kanis model varied, depending on the centre, sex, vertebra, and site concerned. We have seen previously that site-specific thresholds have no advantage over using the same threshold at all sites.

However, the main advantage is likely to be that the predicted posterior height is unbiased. Using the McCloskey-Kanis method, too many vertebrae are reported as having reduced posterior heights, presumably due to the bias observed previously.

11. Comparison of Prevalent Deformity Models

11.1. The Minne Model

The Minne model provided a global model of the vertebral heights. The model was not defined robustly: it had to be defined using a population known to be free from fractures. In order to provide robustness in fitting the model, only the heights of vertebra T4 were used in fitting, since T4 is rarely fractured. This also means that the method is robust to missing data in all vertebrae other than T4.

However, as we saw in Chapter 5, the cubic equation used was not of sufficient order to predict the heights adequately. In addition, the method of fitting, using only the heights of one of the thirteen vertebrae, was inefficient. This method therefore performed poorly, both at predicting heights and at distinguishing between fractured and unfractured vertebrae.

It has some robustness to missing data, since all vertebral heights can be predicted provided that the heights of T4 have been measured. However, if T4 has not

been measured, then *none* of the heights can be predicted. Unfortunately, T4 is missing fairly often, which causes problems for this method.

However, it assumes that all spines are the same shape. The results in Chapters 7 and 8 suggest that this is not the case. This may be part of the reason why this method performed poorly.

11.2. The McCloskey-Kanis Model

This method goes to great lengths to ensure that the models are both defined and fitted in a robust way. The model definition consists of calculating the mean of the ratios H_m/H_p, H_a/H_p and H_p/H_pp for each vertebral level. This means that only the posterior height is used in predicting the other vertebral heights, and so the model definition only needs to be robust against unusual posterior heights.

In order that fractures in adjacent vertebrae did not affect the predicted posterior height, a complex system of trimming was used to elimated any vertebrae with unusually low posterior heights from the calculation of the predicted posterior height. However, we have seen in Chapter 5 that this procedure removed more heights than it should have done, and led to predicted heights that were, in a substantial minority of cases, larger than they should have been.

In addition, deformities in which the posterior height is affected are rare: only 8 out of 301 fractures were classed as crush fractures, in which the posterior height may (or may not) be affected. Therefore, the majority of fractures would not have affected the predicted height and the complex trimming is both unnecessary and

detrimental.

This model is quite robust against missing values. It is possible to evaluate a vertebra provided that all three heights in that vertebra are measured, and at least one of the four adjacent vertebrae. It is not possible to determine whether a height is less than it should be unless it has been measured, so the only additional constraint imposed by this method is that at least one of the 4 adjacent vertebrae be measured in addition.

One advantage of this model is that, unlike the Minne model, it is not a global model. The predicted heights of a given vertebra depend only on the heights of the 4 adjacent vertebrae, not on all of the heights in the spine. Thus it can be applied successfully even if spines are not all the same shape, provided that there is some *local* uniformity of shape.

11.3. The Polynomial Models

The straightforward polynomial model defined in Chapter 5 was not defined or fitted in a robust way. As with the Minne model, we chose a population known to be free from fractures in which to define it. However, a robust way of defining this model was outlined in chapter 6, and shown to to produce very similar results to the simple fitting in subjects free from deformities, but to produce the same model if some of the subjects in the population in which it was defined were given simulated fractures.

The method of fitting the polynomial model outlined in Chapter 5 was not robust to heights that were much less than their predicted values due to fractures. We saw

an example in that chapter of a subject with multiple fractures, in whom all of the predicted heights from the polynomial model were considerably less than the predicted heights from the McCloskey-Kanis model. This led to the robust fitting algorithm outlined in Chapter 6, which did not have this drawback.

This method is extremely robust against missing values. The heights are predicted from the vertebral level, rather than the measured heights. The measured heights are used only to calculate the magnification factor. Therefore, it is possible to predict heights in vertebrae that were not measured using this model, unlike the McCloskey-Kanis model. Only a single vertebra needs to be measured (for the single magnification factor models) in order to predict all vertebral heights.

However, this has the drawback that it is a global model. It was initially assumed that every spine was the same shape, and varied only in size. Eventually, it was made possible to allow the spine to vary in shape (using multiple magnification factors), whilst retaining extremely good robustness to missing values. Provided at least two vertebrae were measured, it is possible to predict all of the heights in the spine, even in the models in which the magnification was allowed to vary. This is slightly better than the McCloskey-Kanis model in which at least one of the four adjacent vertebrae must be measured.

It may be that the restriction to polynomial models was too strict. There are other models that could be used, that provide greater flexibility of shape with fewer parameters. For example, fractional polynomials, as recommended by Royston and Altman [29].

Another alternative would be non-parametric regression. Where linear regression

assumes that the expected height is a linear function of the predictor variables (in our case $H = \sum a_i x^i$), non-parametric regression simply assumes that H is a smooth function of x.

The use of a normal distribution for the error terms may not be appropriate. The long tails in the distribution of residuals shown in Figure 5.1 suggests that a t-distribution may be more appropriate. This is something that could be explored further.

11.4. The Latent Variable Model

The latent variable model of the spine outlined in Chapter 7 was not defined robustly. The latent variables we defined from the covariance matrix of the training sets, which were known to be free from fractures. However, there are methods for obtaining a robust estimate of a covariance matrix, and the latent variable model depends only on the covariance matrix for its definition, so it is possible to define a latent variable model in a robust way.

However, fitting this model robustly is more difficult. The measured heights themselves are used as predictors in this model, so if one of them is much lower than it should be, any heights predicted from it will be much lower than they should be. It may be possible to solve this problem using imputation methods, as used in Chapter 10, but this possibility was not investigated.

Another disadvantage of the latent variable model is its lack of robustness to missing values. Only subjects with measurements for all 39 heights can be used in

this model. Again, it is possible that imputation methods could be used to avoid this disadvantage, but that possibility has not been investigated.

Another criticism that could be levelled at this model is that it does not take into account the structure of the data. The vertebral heights are treated as 39 separate measurements, when in fact they consist of three measurements on each of 13 vertebrae. It may be that a random effects model, that took into account the hierarchical nature of the data, would be more suitable.

11.5. Outlier Detection

The outlier detection method differs from the other methods considered in that it does not produce predicted heights. We are therefore not concerned with the robust fitting of this model.

This method is by definition robust to outliers. Robust estimates of the mean and covariance matrix are used to define the distance of each point from the mean, and the distance from the mean is used to identify outliers.

Missing data is a problem with this method, since the distance from the mean depends on all 39 heights. For subjects with missing data, a statistic was calculated based on the non-missing heights, but this statistic would have a different distribution to that based on all 39 heights. Since the cut-off chosen to determine outliers was based on an approximation to the distribution of the test statistic, it is possible that the approximation works slightly differently depending on the number of heights included, and that subjects with missing values are therefore more or less

likely to be classed as outliers than subjects with all measurements available.

A further disadvantage of this method is that it only identifies *subjects* with deformities, not the deformities themselves. In many situations, identifying subjects is all that is important, but there are situations in which knowing which vertebra is affected would be useful (for example, fractures in the lumbar spine are believed to have more impact on the individual than fractures in the thoracic spine).

Conversely, this method offers a continuous measure of the amount of deformity in the subject, as Minne's Spinal Deformity Index does. This may be of use, for example, in a clinical trial, where it could provide a more accurate measure of the amount of damage in the spine than simply classifying individuals as having a fracture or not having a fracture, or even counting the number of fractures.

11.6. The Imputation Model

The imputation model was based largely on the McCloskey-Kanis model, so it has the advantages of that model. However, it avoids the main disadvantage of that model (the bias in the predicted heights due to overestimating the predicted posterior height).

Only posterior heights are used to predict the other vertebral heights in this model (as in the McCloskey-Kanis model), and as we have seen posterior heights are rarely affected by fractures. Thus robust model definition is not of great importance. However, any posterior heights that are much less than their expected values are replaced by imputed values when defining the model, which does provide a robust

model definition.

The fitting of this model should also be robust, since vertebral heights are predicted only from the posterior heights. Again, any posterior heights which are unusually low (and would therefore lead to unusually low predictions) are replaced by imputed values, and hence the predicted heights should be reliable.

The robustness of this model to missing data is exactly the same as for the McCloskey-Kanis model. The posterior heights of four adjacent vertebrae are used to predict the vertebral heights of any given vertebra, and hence at least one of these adjacent vertebrae must be measured in order to obtain prediction. However, fracture status can only be determined in those vertebrae which were actually measured, since it is determined by comparing the measured and predicted heights.

This model offered the best performance in defining prevalent vertebral deformities. It is conceptually and computationally very similar to the McCloskey-Kanis method, which is already widely used. It would therefore make a very good direct replacement for this method.

Part III.

Incident Fractures

12. Introduction to Incident Vertebral Deformities

Incident deformities are those which are known to have occurred during a particular time interval. In order to identify such deformities, at least two x-rays are required: the first one in which there is no evidence of a deformity and the second one in which the deformity is apparent.

Incident deformities are generally of more interest than prevalent deformities. This is because with a prevalent deformity, it is not known when the deformity occurred. It could have happened when the individual was a child, or it could have happened very recently. This makes it difficult to interpret associations with risk factors: for risk factors that change over time, it is impossible to say whether the risk factor was present at the time the deformity occurred.

This is particularly important in clinical trials. When a subject is recruited to a trial of an anti-osteoporotic drug, a baseline x-ray will be taken to determine if any fractures are already present. However, the outcome of interest in the trial is the

number of *new* fractures that occur after treatment has been started.

There are a number of important ways in which identifying incident deformities differs from identifying prevalent deformities. Firstly, there is twice as much information available, since there are two x-rays. In addition, we are looking for heights that are unusual compared to their previous measurement, rather than just unusual compared to the expected values in the population. This should make incident deformities easier to identify.

Another factor helping to make incident deformities easier to identify is the fact that any large change in the comparatively short period between x-rays is unlikely to be due to any reason other than a fracture. Congenital deformations exist throughout life and tend not to change, and even degenerative disease (osteoarthritis) proceeds very slowly.

On the other hand, there will be far fewer incident deformities than prevalent deformities, since any deformity that occurred between birth and the baseline film (a period of between 50 and 80 years) will be classed as a prevalent deformity whilst only deformities that occurred during the followup (a period of around 4 years) will be classed as incident deformities. Thus, in the three centres in which we investigate prevalent deformities, 140 subjects had prevalent deformities but only 37 had incident deformities.

This has important consequences for the properties of a method of identifying incident deformities. The McCloskey-Kanis algorithm used to define prevalent deformities had a false positive rate of around 5%, and the prevalence of vertebral deformity was around 15%. Thus, 75% of the subjects classed as having a deformity

had a genuine fracture.

However, since less than 5% of subjects had one or more incident deformities, using a method with the same false positive rate would mean that fewer than 50% of the subjects classed as having deformities would have a genuine fracture. Therefore, the method used to define incident deformities must be more specific than that used to define prevalent deformities. Some details on how choices of sensitivity and specificity affect the power to detect associations between risk factors and vertebral deformities are given in Chapter 14.

Chapter 14 also outlines the two main approaches taken to defining incident deformities in the literature, and proposes a third method which is a combination of these two. The methods are compared using not just the clinical opinion as a referent (since this is not universally accepted as a gold standard: two clinicians may disagree about the presence of a fracture), but also a number of risk factors for vertebral fracture.

One of the main problems with using 2 radiographs to identify deformities is the fact that the magnification may differ between the two films. If the magnification of the second film is less than the first, the heights will all appear to be reduced, which may lead to false positives (depending on the method of defining incident deformities). Ideally, the magnification factors of the films will be recorded as described in Chapter 2, but this was not always done in the EVOS study. Therefore, a statistical method of adjusting for possible magnification differences was developed, and is explained in Chapter 13.

The results of fitting a number of established models are given in Chapter 15.

The performance of a method based on the robust polynomial models of Chapter 6 is given in Chapter 16 and that of a method based on the *ad hoc* method of Chapter 10 is given in Chapter 17.Finally, the different models are compared in Chapter 18.

No attempt was made to extend the latent variable models of Chapter 7 or the outlier methods of Chapter 9 to define incident deformities. The latent variable models were difficult to make robust, and did not perform particularly well even for identifying prevalent deformities. In addition, considering incident deformities doubles the dimensionality of the space to be considered (from 39 heights to 78 heights on two films), and the numbers of subjects available in individual centres was often less than this. This would mean that the covariance matrix would be singular.

13. Magnification Differences Between Consecutive Radiographs

13.1. Introduction

The work presented in this chapter was performed for the EPOS study to examine whether we could correct for differences in magnification between consecutive radiographs when the spine-film distance had not been recorded. A pilot study using only the data from a single centre, Malmo, was performed. It has not been possible to extend this analysis to other centres since the spine-film distances were not made available for these other centres.

13.2. Methods

13.2.1. Magnification of Radiographs

The data set on which we tested our methods consisted of paired radiographs, taken 2 years apart, on 423 subjects from Malmo. For all subjects, the film-focus distance (F) was fixed at 120cm for both visits, and no changes in equipment were made between the two visits. If the spine-film distance (S) is known, the magnification of the radiograph can be calculated as $\frac{F}{F-S}$, so that the heights measured on the x-ray film can be converted to the true heights by using

$$corrected\ height\ =\ \frac{measured\ height \times (F-S)}{F} \tag{13.1}$$

Using the above equation to correct for differences in magnification will be referred to as "theoretical" correction, to distinguish it from our new "empirical" method. The magnification of the second film relative to the first is

$$magnification\ =\ \frac{(F-S_2)}{(F-S_1)} \tag{13.2}$$

since F did not change between the two films.

For 112 of these subjects, the spine-film distance was known for both films, and could be used to adjust for magnification. The mean magnification factor for these subjects was 1.3. For the remaining 311 subjects, the spine-film distance was only recorded for 1 of the two radiographs. For these subjects, the raw heights, as measured directly from the radiograph, were used.

13.3. Statistical Methods

For each subject, the ratio of each of the 39 heights per examination on the second film to the corresponding height on the first film was calculated. If no new deformity had occurred, the mean of all of these ratios could be taken as the magnification of the second film relative to the first film. By multiplying the heights on the first film by this scaling factor, any differences in magnification should in principle be corrected.

However, if a fracture occurred in one or more vertebrae between the two films, or one of the points on either film were misplaced, this large difference between the films would bias the estimated scaling factor. To adjust for this, a trimmed mean was used for the scaling factor, using the method suggested by Melton [30] to remove outliers (magnification ratios that fell more than 1.5 interquartile ranges beyond the 25th and 75th percentiles). Heights that have been adjusted by an empirical scale factor in this way are referred to as 'empirically corrected' heights, to differentiate them from the 'theoretically corrected' heights obtained conventionally using scale factors calculated from the film-focus and spine-film distances.

In practice, each x-ray consists of two films, one of the lumbar spine and one of the thoracic spine. These two films may be at slightly different magnifications. However, the vertebrae from T10-L1 may appear on either or both films, and were read from whichever film had the clearer image. Thus, for a given individual, there may be four different magnification factors:

1. Vertebra read from thoracic film on both occasions.

2. Vertebra read from thoracic film on first occasion, lumbar film on second occasion.

3. Vertebra read from lumbar film on first occasion, thoracic film on second occasion.

4. Vertebra read from lumbar film on both occasions.

T4-T9 were always in group 1, and L1-L4 were always in group 4. However, T10-T12 could appear in any of the four groups, and which group they actually belonged to was not recorded. Thus correcting the magnification for each film separately was not possible. However, we looked at the effect of fitting separate magnification factors to the upper 4 and lower 4 vertebrae (T4–T7 and L1–L4) to see if adjusting each film separately might produce a better fit if it were possible.

13.4. Results

13.4.1. Agreement Between Theoretical and Empirical Correction

The spine-film distance was recorded to the nearest cm. On the first round, spine-film distances ranged from 26cm to 31cm, whilst on the second film these distances ranged from 17cm to 29cm. The correlation between the spine film distances on the first and second round was strong (r=0.44), but the second round distance was almost always less than the first round distance: see table 13.1.

The empirical correction factors had a mean of 0.98 and a standard deviation of 0.015, whilst the theoretical correction factors (in these subjects) had a mean of 0.98

Change in Distance	Number of Subjects
-11 cm	1
-5 cm	3
-4 cm	10
-3 cm	33
-2 cm	35
-1 cm	19
0 cm	8
+1 cm	1
Total	110

Table 13.1.: Change in spine-film distance between the x-rays

and and standard deviation of 0.017. Figure 13.1 shows the relationship between the empirical and theoretical correction factors for each subject. It can be seen that the theoretical correction factor only takes a certain number of fixed values, whilst the empirical correction factor can take any value. The correlation between the two correction factors is statistically significant but not strong ($r = 0.23$, $p = 0.03$).

13.4.2. Changes in Height between Films

In the 84 subjects with no clinical abnormalities and with the spine-film distance recorded for the second film, the ratio of each height on the second film to the corresponding height on the first film had a mean value of 0.98 and a standard deviation of 0.050. 13.7% of the observed variation was due to systematic differences

between subjects. Fitting the theoretical correction factor for each subject accounted for only 0.5% of the total variation, or 4.0% of the between-subject variation. Fitting the empirical correction factor accounted for all of the between subject variation, and reduced the within-subject standard deviation to 0.047.

Figure 13.2 shows the distribution of relative changes in height in the 84 clinically normal subjects with known spine-film distances. Both methods of correction move the distribution so that it is centred on 0% rather than -2.5% (i.e. they remove the bias caused by the fact that the second films were generally at a slightly smaller magnification than the first films). The distribution of empirically corrected changes has a higher peak and is less spread out than the distribution of theoretically corrected changes, demonstrating better measurement precision.

Figure 13.3 shows the relative change in height between the two films for each vertebral height in all 316 subjects without clinical abnormalities. It can be seen that the empirical correction moves the entire distribution to the right to be centred on 0% and makes the distribution more 'peaked'.

Fitting two separate magnification factors to each subject accounted for a further 9.1% of the observed variation, a highly significant improvement in fit ($p < 0.0001$). The within subject standard deviation was reduced to 0.045.

13.4.3. Effect on False Positive Rate

There were 3246 heights measured in 84 subjects with a spine-film distance recorded, and no clinical abnormality. If no correction for magnification was applied, 30 (0.9%) showed a decrease of 15% or more. Applying the theoretical correction reduced the

number to 11 (0.3%), whilst applying the empirical correction reduced the number to 14 (0.4%). Both reductions were statistically significant using McNemar's test to allow for the matching, but not significantly different from each other. The number of heights showing a reduction of 20% or more was reduced from 3 to 2 whichever method of magnification correction was used, but this change was not statistically significant.

In the 232 subjects without spine-film distances recorded for the second film and no clinical abnormality, 8808 heights were measured. If no magnification correction was applied, 70 (0.79%) showed a reduction of 15% or more and 12 (0.14%) showed a reduction of 20% or more. Applying the empirical magnification correction reduced the numbers of heights showing large reductions to 35 (0.40%) and 7 (0.08%) respectively. Both reductions were statistically significant using McNemar's test.

13.4.4. Effect on False Negative Rate

When the films were assessed by a radiologist, 18 vertebrae were found to have suffered an incident fracture: 15 amongst subjects with unknown magnification and 3 among subjects with known magnification. All 18 changed by 20% or more, irrespective of which method of correction for magnification was applied, or indeed if none was applied. Thus the false negative rate in this group of subjects was zero, and unaffected by correction for magnification.

13.5. Discussion

We have shown that the magnification of serial radiographs of an individual subject may differ, even if all possible steps to avoid it are taken. If the change between two radiographs is to be used as a criterion for detecting incident deformities, changes due to differences in magnification must be eliminated. In our dataset, the magnification of the second film generally less than the magnification of the first film, which if left uncorrected would lead to a greater than necessary number of false positives.

The standard method of removing differences in magnification is to measure the spine film distance and calculate the expected magnification from that and the film-focus distance. We have developed a method that can be used even if the spine film distance was not recorded. It reduces the imprecision of the vertebral height measurements more than correction using the spine-film distance. It also significantly reduces the number of vertebrae incorrectly classed as incident deformities. In this, it performed neither better nor worse than the conventional method.

Each radiograph consists of two separate films, a lumbar and a thoracic film. In theory, these two films may be at different magnifications, and hence separate magnification factors should be fitted to them. Since the films overlap, there are certain vertebrae that we could not be certain from which film they had been read. Fitting two magnification factors to those heights that could only be read from one film did reduce the measurement error markedly, but we could not determine whether this lead to a significant decrease in false positive deformities.

In conclusion, we have developed a method of correcting for differences in mag-

nification between consecutive radiographs of the same subject, that does not require measurement of the spine film distance. It improves the precision of vertebral height measurements and reduces the number of false positive incident deformities by between 1/3 and 1/2, without significantly affecting sensitivity. Application of this technique in cohort studies and clinical trials with vertebral fracture as an end point should considerably enhance the statistical power of such studies, or reduce the demands of quality assurance, or both.

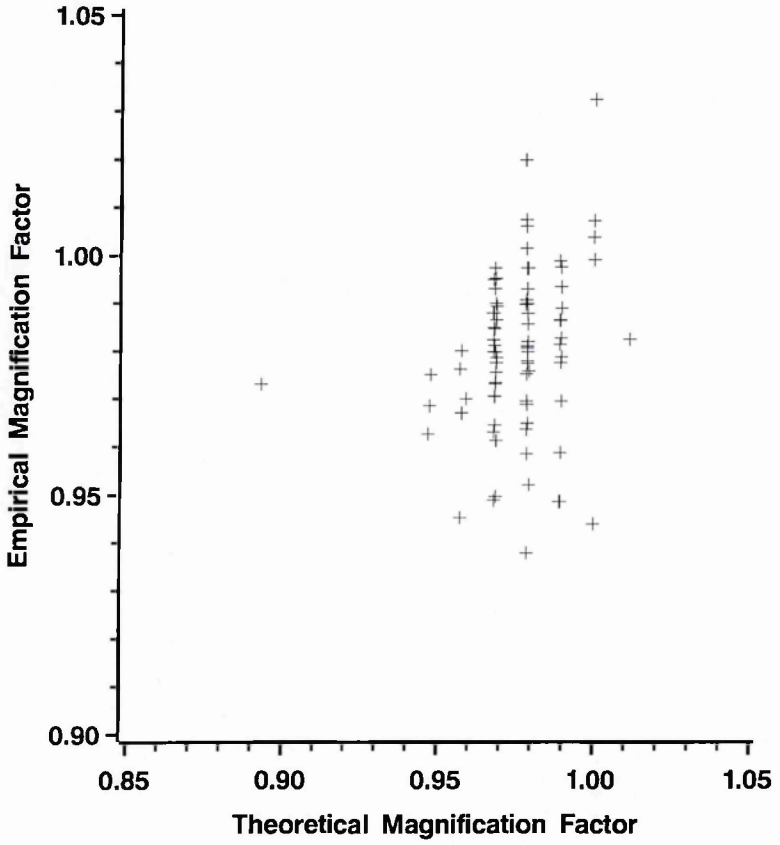


Figure 13.1.: Relationship between theoretical and empirical magnification factors in 84 subjects with no clinical abnormality.

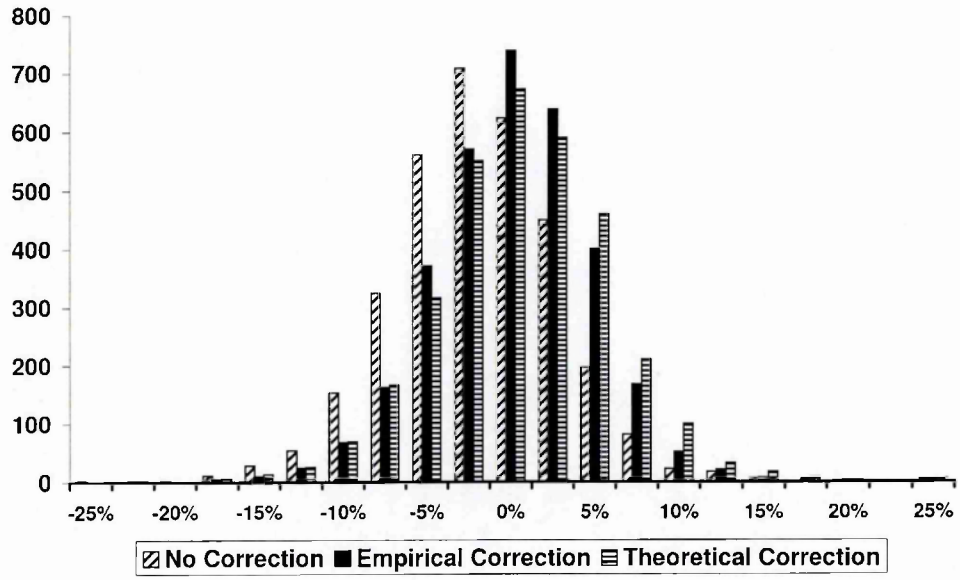


Figure 13.2.: Distribution of relative change vertebral heights in 84 subjects with no clinical abnormality and known spine-film distance

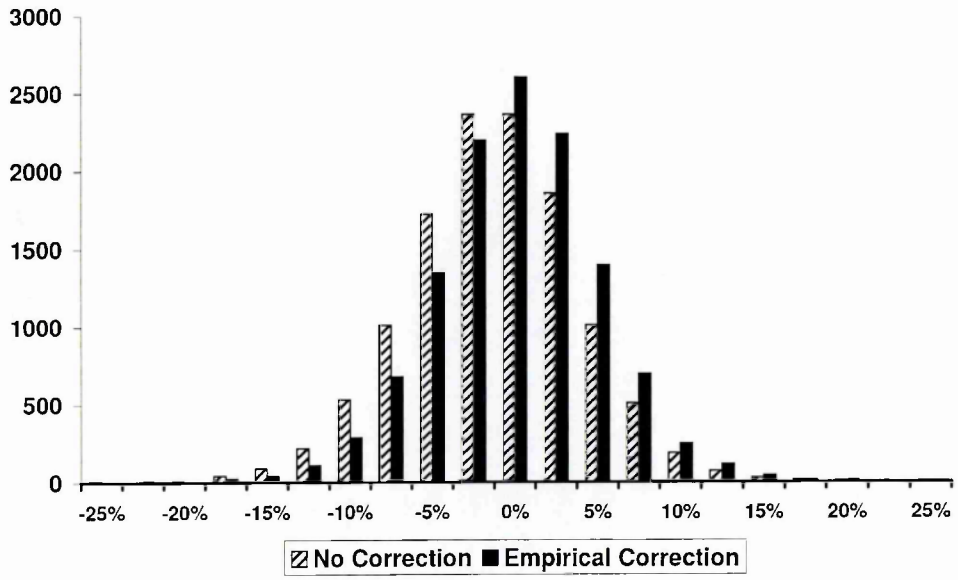


Figure 13.3.: Distribution of relative change vertebral heights in 316 subjects with no clinical abnormality

14. Approaches to Defining Incident Deformities

Methods of defining incident vertebral deformities aim to identify a change of state in the vertebrae studied during a defined time period. There are fundamentally two different approaches to defining an incident vertebral deformity given two consecutive radiographs. One approach is to consider the shape of each vertebra on both films: if it is a normal shape on the first film and an abnormal shape on the second film, it is classed as an incident deformity. A number of methods of defining an abnormal shape have been proposed [8, 9, 10, 11]

This approach is commonly referred to as the *point prevalence* approach and has the advantage that the same definition can be used for incident and prevalent deformities. Also, since only shape is considered, the method is unaffected by changes in magnification between the two radiographs. However, the sensitivity and specificity of this approach depends on the distribution of normal shapes in the population concerned: if the range is narrow, then deformities will be easy to detect but if it is wide, this will be harder. Furthermore, it cannot detect further deformation in a

pre-existing deformity.

The alternative is to compare the vertebral dimensions, in this case heights, on the second film to those on the first film. If any height has reduced beyond a certain threshold, the vertebra is classed as having an incident deformity. A commonly used threshold is a reduction of both 20% and 4mm [31], although others have been proposed [32].

This approach has the advantage that the sensitivity and specificity do not depend on the distribution of vertebral shapes in the population, only on the measurement error. However, this method is unreliable if the magnifications of the two radiographs are different. Furthermore, a vertebra can be classed as an incident deformity by this method without necessarily being classed as a prevalent deformity by any method.

When comparing methods of defining incident deformities, it is important to consider the reason for doing so, since this will influence which method is considered better. If the purpose were to diagnose fractures in an individual, in order to help decide whether treatment to prevent further bone loss is necessary, then a very sensitive method is required. Specificity is less of a concern, since giving an unnecessary, but fairly benign treatment is less serious than failing to give a needed treatment. However, the concern here is with developing a suitable definition for use as an endpoint in clinical trials and epidemiological studies. We are concerned with obtaining accurate and precise estimates of population level parameters (relative risks, treatment effects etc.) and the effect of an imperfect method of defining fractures can be quantified as the bias or loss of precision in such estimates.

A number of methods have been compared by Black et al. [33], also thinking primarily of an epidemiological end-point. They compared point prevalence methods based on definitions of vertebral deformity proposed by Minne [6], McCloskey[11] and Melton and Eastell [8, 9] with a modification by Black [10]. They also considered a straightforward percentage reduction in height[32], and a number of slight modifications to this method. They compared the proportion of subjects with a given risk factor or outcome measure among those classed as having incident deformities to the proportion among those classed as being without deformities, and concluded that no method was markedly better than another.

It seems reasonable to require incident deformities to *both* be an abnormal shape on the second radiograph and to have changed appreciably since the initial radiograph. In addition, requiring a vertebra to satisfy two conditions rather than one may improve specificity (fewer vertebrae will satisfy both conditions by chance than will satisfy either one), and if incident deformities are rare, specificity will be more important than sensitivity in determining the performance of the method, as we shall show.

We hypothesised that combining both approaches would give a better definition of incident vertebral deformity. We compared the different definitions by assessing their agreement with

1. A clinical opinion as to the presence of an incident vertebral deformity.

2. Risk of vertebral fracture calculated from a number of risk factors for vertebral deformity measured at baseline (age, gender, bone mineral density, prevalence

of deformities in other vertebrae).

14.1. Methods

14.1.1. Predictors

Bone Density Measurements

Bone mineral density was measured using Dual Energy X-ray Absorptiometry (DXA). This method consists of passing X-rays though the body, and creating a two dimensional image of the bone as it passes though. By using two different frequencies of X-ray and comparing the the absorption of each, it is possible to determine mass of bone that the X-rays have passed through. A computerised edge-detection algorithm then determines the area of bone on the two dimensional image, and by dividing mass by area it gives and areal density. This areal density has been shown to be very strongly associated with the risk of fracture. Although there are methods to measure true volumetric bone densities *in vivo*, they are not widely used clinically since they do not offer sufficient improvement in the ability to predict fracture to justify the additional cost.

The densitometers in each centre were, with one exception (a Sopha fan-beam machine), pencil beam machines made by Lunar, Hologic or Norland and were cross-calibrated using the European Spine Phantom prototype [34, 35, 36]. This is a semi-anthropomorphic phantom with three 'vertebrae' of specified densities 0.5g/cm^2, 1.0g/cm^2 and 1.5g/cm^2. At least 5 measurements of the phantom were made on each machine and a two-parameter empirically fitted exponential calibration curve

used to convert measured density values into standardised values, as described by Pearson et al.[36]. Detailed descriptions of the densitometry procedures are to be found in the paper by Lunt et al. [37]

14.1.2. Outcome Measurements

New Prevalent Deformity Method

There are a number of different approaches to defining prevalent deformities [6, 8, 9, 38, 11, 10], and any of these can be used for the new prevalent deformity approach. We chose to use the McCloskey-Kanis algorithm [11] to identify prevalent deformities, since it has been shown to work at least as well as any of the others[10], and had already been used to define prevalent deformities in the first phase of this study.

The Height Reduction Method

One widely used definition of incident vertebral deformity [39, 31] is based on the amount of change in height between two X-rays. If any of the three vertebral heights have decreased by 20% or more, and the decrease is at least 4mm, the vertebra is considered to be an incident deformity.

A Combination Method

We also considered a definition of incident deformity in which a vertebra had to both

1. be classified as a McCloskey-Kanis deformity on the second x-ray; *and*

2. one of the heights in the vertebra must have reduced by at least 20% and by at least 4mm.

Such deformities were subdivided according to whether they were already classed as being deformed on the first x-ray.

14.1.3. Statistical Methods

Two approaches were used to compare the different methods of defining incident fractures, discriminant analysis and calculating the expected bias and efficiency of each method. The reason for this was that there are different groups who need to be convinced of the superiority of a given method, epidemiologists and radiologists. The epidemiologists mistrust a subjective measure, such as a clinical opinion, and would not accept it being used as a gold standard, so for this group we used discriminant analysis in which clinical opinion was used as a predictor of fracture, but not as a gold standard. The radiologists, on the other hand, do believe that a clinical opinion should be used as a gold standard (despite the fact that each radiologist will provide a slightly different gold standard). Measuring the expected bias and efficiency tests which method gives the best agreement with the individual radiologist who read these films, which may not be a gold standard but should agree well with it.

Discriminant Analysis

Discriminant analysis [40] is a statistical technique to allocate subjects to one of two groups based on a number of predictor variables. Given predictors $x_1, x_2, \ldots x_p$, a

linear function of of the predictor variables

$$Z = \sum_{i=1}^{p} a_i x_i$$

is calculated. The optimal choice of Z is that function which, if a t-test to compare the scores in the two groups were performed, the t-statistic would have the largest magnitude. The groups are then "further apart" in this direction of a p-dimensional space than in any other direction.

There are an infinite number of suitable choices for the coefficients a_i, since multiplying all of the coefficients by the same constant gives and equivalent discriminant function. Values for the a_i can be calculated by performing linear regression, using the x_i as predictors and an outcome variable Y which is determined by which group an individual belongs to. The choice of values given to Y is arbitrary: different values will give different, but equivalent, discriminant functions. One common choice is to use 0 and 100 for the two values of Y, and that is what we did here.

We developed a discriminant function using age, gender, bone mineral density (BMD) and the clinical opinion as to the presence of an incident deformity to distinguish between the subjects whom all three quantitative methods agreed had incident deformities and the subjects whom all three quantitative methods agreed had no incident deformities. We then applied that function to the subjects about whom the methods disagreed to see whether they were more similar to the agreed cases or the agreed non-cases.

Six separate discrimination functions were produced, containing

1. Age, sex and baseline prevalent deformity

2. Age, sex, baseline prevalent deformity and spine BMD

3. Age, sex, baseline prevalent deformity and hip BMD

4. Age, sex, baseline prevalent deformity and the clinical opinion.

5. Age, sex, baseline prevalent deformity, clinical opinion and spine BMD.

6. Age, sex, baseline prevalent deformity, clinical opinion and hip BMD.

Effect of Choice of Morphometric Method on Study Power

If an imperfect measure is used to define cases in a study, some cases are likely to be misclassified as non-cases and vice versa. This will lead to a bias in the estimate of the association between risk factors and the outcome, and a loss of efficiency (i.e. more subjects are required to achieve the same level of significance using an imperfect measure than using a perfect measure). The bias and loss of efficiency depend on the sensitivity and specificity of the outcome measurement, the proportion of the population who are cases and the distribution of the risk factor in cases and non-cases.

We considered a dichotomous risk factor, since this is the simplest case. We calculated the bias and loss of efficiency that could be expected with each method, allowing the prevalence of the risk factor and its odds ratio to vary. The details of these calculations are given in the next two sections.

14.1.4. Calculation of Bias in the Odds Ratio due to Misclassification of Outcome in a 2x2 Table

Suppose that the *true* distribution of outcome and dichotomous exposure variables in a group of subjects in a study is as in the table below.

	Exposed	Unexposed
Case	a	b
Non-Case	c	d

Then if a test with sensitivity Se and specificity Sp is used to classify subjects as either cases or non-cases, some subjects will be classified incorrectly. The proportion of cases correctly classified will be Se, whilst a proportion $(1 - Se)$ of cases will be incorrectly classified as non-cases (this is the definition of sensitivity). Equally, Sp of the non-cases will be correctly classified as non-cases, but $(1-Sp)$ will be incorrectly classified as cases. If we assume that there is no misclassification of exposure status, and that the misclassification of case status does not differ between exposed and non-exposed subjects, then the 2 x 2 table that would result from the study would be:

	Exposed	Unexposed
Case	$a \times Se + c \times (1 - Sp)$	$b \times Se + d \times (1 - Sp)$
Non-Case	$c \times Sp + a \times (1 - Se)$	$d \times Sp + b \times (1 - Se)$

Hence the observed value of the odds ratio will not be $OR_{pop} = ad/bc$ (the true value in the study subjects) but

$$OR_{obs} = \frac{\{a \times Se + c \times (1 - Sp)\} \times \{d \times Sp + b \times (1 - Se)\}}{\{b \times Se + d \times (1 - Sp)\} \times \{c \times Sp + a \times (1 - Se)\}}$$

The extent to which OR_{obs} differs from OR_{pop} is a measure of the bias in the estimation of the odds ratio.

14.1.5. Calculation of Loss of Efficiency Due to Misclassification of Outcome in a 2x2 Table

Suppose that the proportions of subjects in each of the cells of the 2 x 2 table are:

	Exposed	Unexposed
Case	p_a	p_b
Non-Case	p_c	p_d

Then the χ^2 statistic is given by

$$\begin{aligned}
\chi^2 &= \frac{((p_a N \times p_d N) - (p_b N \times p_c N))^2 N}{(p_a N + p_b N) \times (p_c N + p_d N) \times (p_a N + p_c N) \times (p_b N + p_d N)} \\
&= \frac{((p_a \times p_d) - (p_b \times p_c))^2}{(p_a + p_b) \times (p_c + p_d) \times (p_a + p_c) \times (p_b + p_d)} \times N \\
&= kN
\end{aligned}$$

where k depends only on the proportions of subjects in each cell, not the absolute numbers. In other words, if we multiply the number of subjects in each cell of the 2 x 2 table by any factor, the χ^2 statistic increases by the same factor.

So we can write $\chi^2_{hyp} = k_{hyp}N$ and $\chi^2_{obs} = k_{obs}N$.

Now, the efficiency of a study is defined as

$$\text{Efficiency} = \frac{N_{hyp}}{N_{obs}}$$

where N_{obs} is the number of subjects required using the imperfect diagnostic tool to achieve the same power as a study of N_{hyp} subjects with a perfect diagnostic tool. Since equal power means that the χ^2 statistics are equal, we have

$$\begin{aligned} k_{hyp}N_{hyp} &= k_{obs}N_{obs} \\ \Rightarrow \quad \frac{N_{hyp}}{N_{obs}} &= \frac{k_{obs}}{k_{hyp}} \end{aligned}$$

So the efficiency is equal to $\frac{k_{obs}}{k_{hyp}}$.

Now, if the studies were performed with the same of subjects, N, in each study, the χ^2 statistics for the two studies would be Nk_{hyp} and Nk_{obs}, and their ratio would be $\frac{k_{obs}}{k_{hyp}}$.

I.e. the efficiency is given by the ratio of the χ^2 statistics in studies with the same number of subjects.

14.2. Results

14.2.1. Distribution of Predictors

Table 14.1 gives the distribution of the variables used in the discriminant analysis,

% Female	53.2
% With Prevalent Deformity	10.4
% With Qualitative Incident Deformity	3.6
Mean Age in years (SD)	62.6 (7.7)
Mean Spine BMD in g/cm^2 (SD)	0.99 (0.22)
Mean Trochanteric BMD in g/cm^2 (SD)	0.78 (0.15)
Mean Femoral Neck BMD in g/cm^2 (SD)	0.69 (0.15)

Table 14.1.: Distribution of discriminant variables

14.2.2. Agreement Between Morphometric Definitions

A total of 77023 vertebrae were evaluated, of which 640 were classed as incident deformities by at least one method. There were 276 vertebrae that were deformities using all three methods, 286 that were positive by the point prevalence method but not the others, and 44 that were positive by the height reduction method but not the others. In addition, there were 34 vertebrae that were positive by the combination and height reduction methods, but not the point prevalence method. These were vertebrae that were classed as prevalent deformities at baseline, but which showed a marked reduction in at least one height during the follow-up period.

Since the predictor variables were measured on subjects, not on individual vertebrae, the analysis had to be performed on subjects. Figure 14.1 shows the agreement between the point prevalence and height reduction methods at the subject level. The numbers in brackets give the number of subjects considered to have an incident deformity by the radiologist.

Whilst it is impossible for a *vertebra* to satisfy both the height reduction and point prevalence definitions without satisfying the combination definition, this is not true of subjects, if the different definitions identify different vertebrae as incident deformities. This occurred in one subject, who was not included in figure 14.1, nor in the subsequent analysis.

In addition, there were 7 subjects who satisfied both the height reduction and combination definitions, but not the point prevalence definition. Ideally, this group would be analysed separately, but since it is so small, it was also excluded from the analysis.

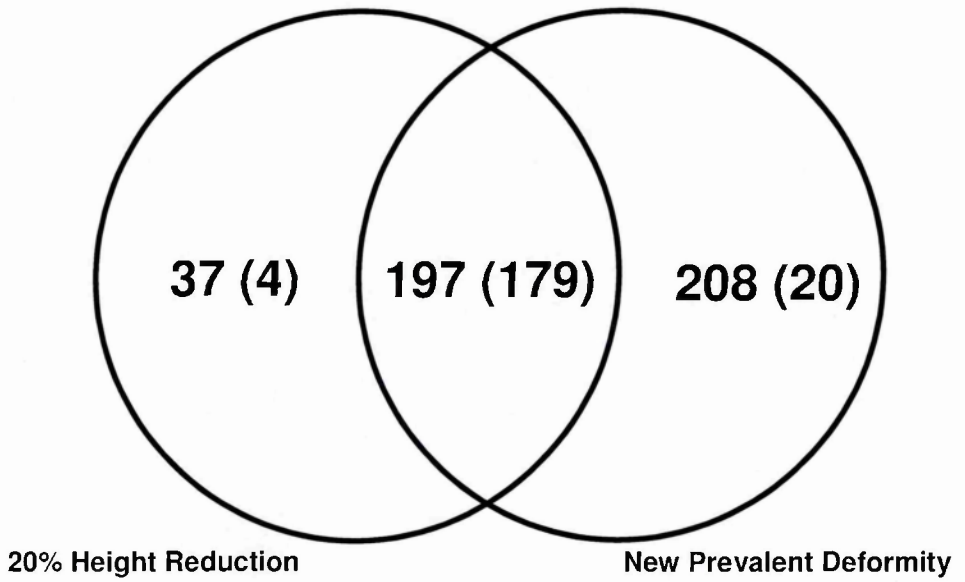


Figure 14.1.: Agreement between morphometric definitions of deformity in subjects

14.2.3. Discriminant Analysis

The mean value of the discriminant function for each of the groups for the six discriminant models are given in table 14.2. They also show the size of each group.

Although there are highly significant differences between groups using all discriminant functions, it is clear that the discrimination is much better (the score in the "all three" group is closer to its theoretical ideal value of 100) if the qualitative evaluation is included. In the three discriminant functions including the qualitative evaluation, the subjects classed as having incident deformities by either the point prevalence method (but not the height reduction method) or the height reduction method (but not the point prevalence method) are more similar to the subjects agreed to be normal than the subjects that satisfy all three definitions.

Variables in Discriminant Function	Mean Discriminant Score (Standard Deviation) in Subjects With			
	No Incident Deformity	Height Reduction Only	Point Prevalence Only	Combination
	(N = 6344)	(N = 37)	(N = 208)	(N = 197)
Basic	2.92 (3.06)	4.08 (4.16)	5.10 (4.22)	6.34 (4.62)
Basic & Spine BMD	2.74 (3.17)	3.61 (2.78)	4.70 (4.02)	6.59 (4.24)
Basic & Hip BMD	2.59 (3.24)	4.13 (3.68)	4.18 (4.27)	6.88 (4.79)
Basic & Qualitative Opinion	0.81 (6.51)	9.26 (25.51)	8.55 (23.78)	74.14 (23.21)
Basic & Qualitative Opinion & Spine BMD	0.82 (6.68)	0.49 (0.71)	4.39 (16.86)	72.19 (21.71)
Basic & Qualitative Opinion & Hip BMD	0.53 (5.02)	11.41 (30.09)	5.10 (19.52)	80.89 (25.25)

Table 14.2.: Discrimination between morphometric methods using different discriminant functions

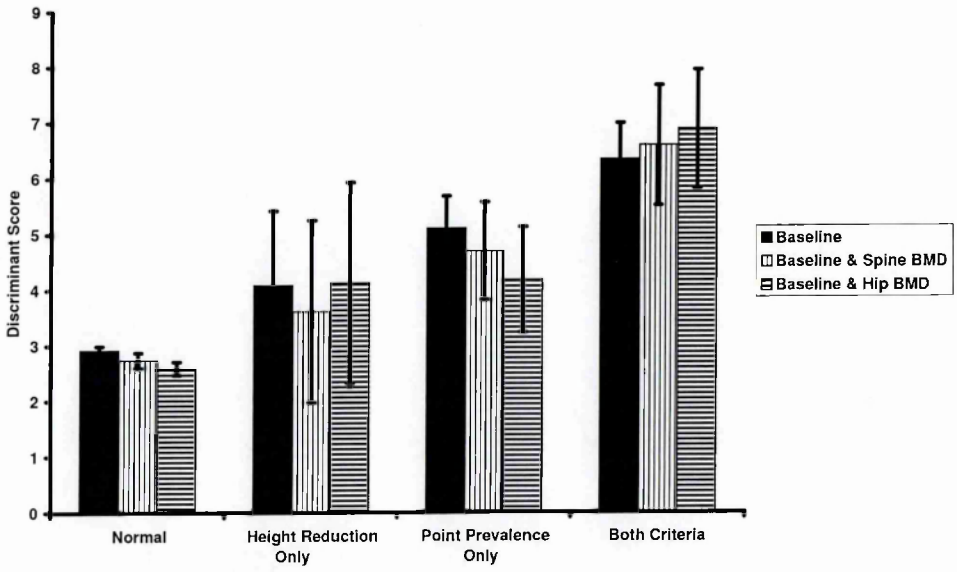


Figure 14.2.: Discriminant scores, excluding qualitative evaluation, in each morphometric group

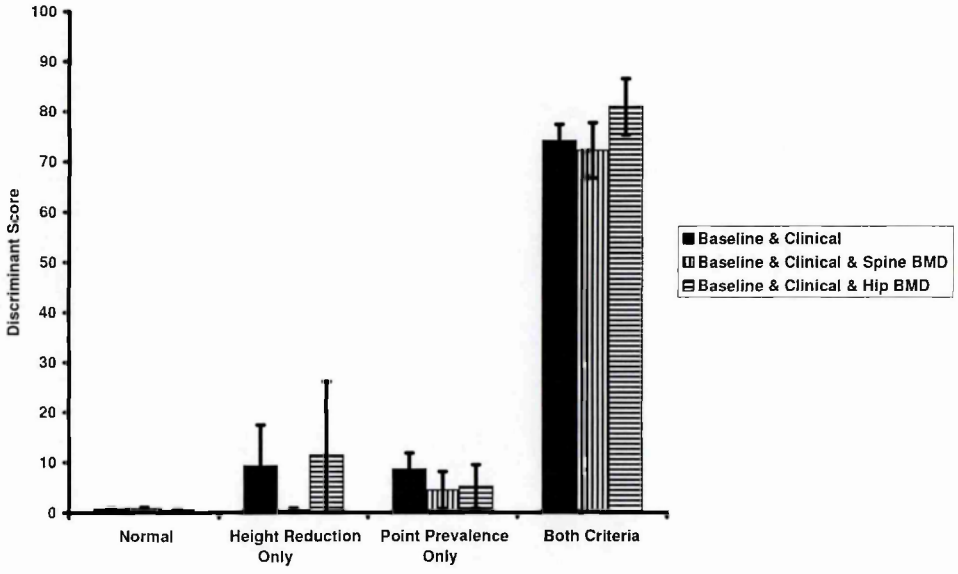


Figure 14.3.: Discriminant scores, including qualitative evaluation, in each morphometric group

14.2.4. Effect of Choice of Morphometric Method on Study Power and Bias

Using the qualitative evaluation as the gold standard, the sensitivity and specificity of the three methods were estimated as

- Point Prevalence: Sensitivity = 81.3%, Specificity = 96.8%
- Height Reduction: Sensitivity = 76.0%, Specificity = 99.1%
- Combination: Sensitivity = 74.4%, Specificity = 99.6%

Figures 14.4 and 14.5 show how the estimated odds ratio of a hypothetical risk factor and the efficiency of a hypothetical study would vary with the proportion of subjects who are genuine cases. For these figures, it was assumed that 50% of the population were exposed and the true odds ratio was 2, although changing these parameters did not affect the shape of the graphs.

It can be seen that the combination method is the most efficient if the proportion of cases is small, but differs very little from the height reduction method if more than approximately 15% of the population have suffered incident events. The combination method shows less bias in the estimation of the odds ratio if less than approximately 40% of the population have suffered incident events.

14.3. Discussion

The combination method proposed in this paper shows a number of advantages over the single criterion methods. Not only does it show better agreement with the

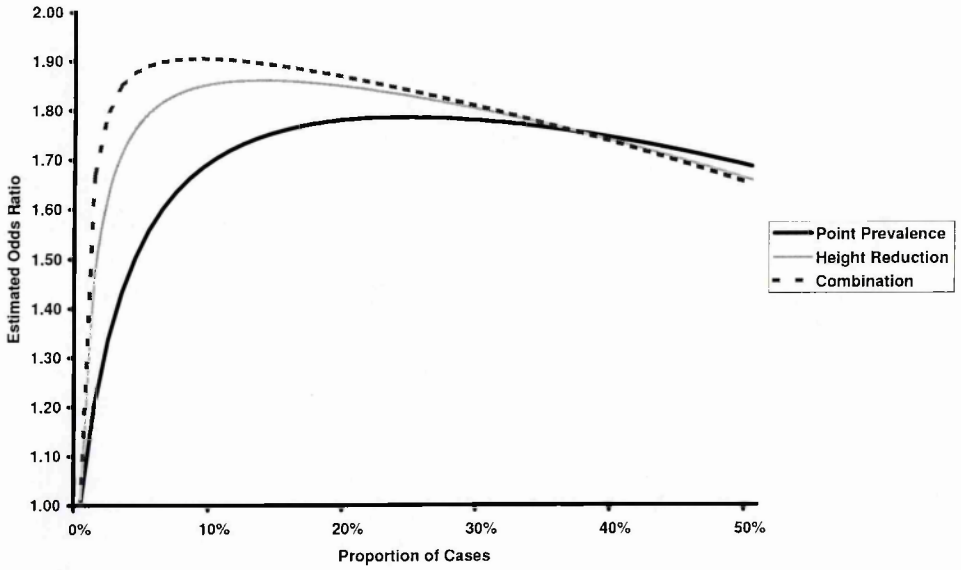


Figure 14.4.: Effect of proportion of population with incident fractures on estimated odds ratio using 3 morphometric methods

qualitative evaluation of an experienced radiologist and stronger association with known risk factors, but it offers greater statistical power in a study of a given size, and less bias in the estimation of the effect of a risk factor, provided that incident deformities are comparatively rare.

There are a number of limitations to this study. For the purposes of calculating the bias and loss of efficiency, the sensitivity and specificity of each method was calculated based on a single radiological opinion, not a gold standard. However, the differences in bias and efficiency of the different methods depend on the differences in sensitivity and specificity between the different methods, rather than their absolute

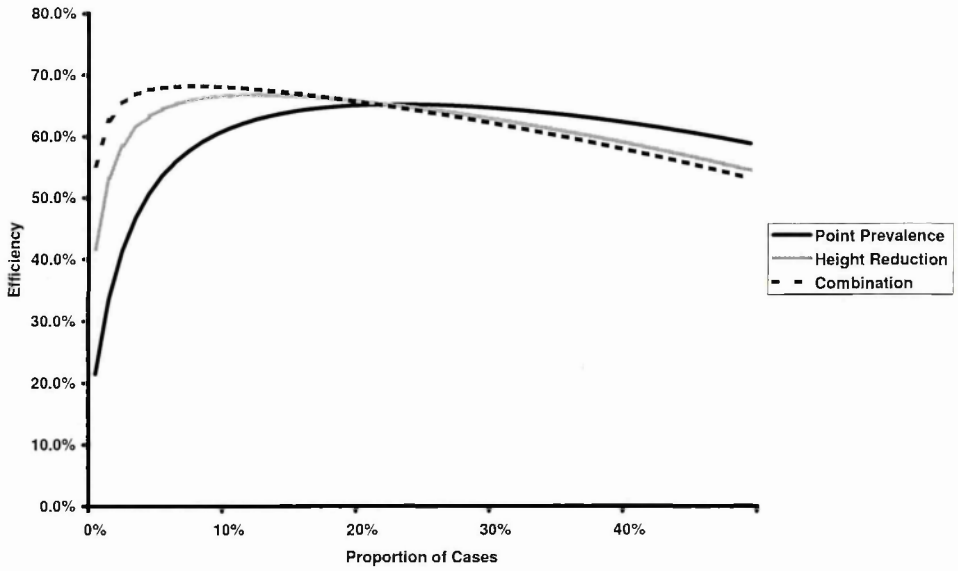


Figure 14.5.: Effect of proportion of population with incident fractures on efficiency of study using 3 morphometric methods

levels. Thus, it is only the subjects for which the methods were discordant that are important, and this is a small number.

Only certain biological correlates were included in the discriminant analysis. However, they were the variables most strongly associated with incident vertebral deformities in this cohort, and other potential risk factors showed only weak or non-significant associations [41].

BMD was only available in a subset of subjects, and this was not invariably measured at baseline. It was, however, measured soon after baseline in most cases, and thus the measured value is likely to be little different from the value at baseline.

At the spine, an incident fracture at L2-L4 could lead to an abnormally high BMD at that site, but this would tend to reduce the association between low BMD and incident vertebral deformity and thus reduce the power of the discriminant analysis. However, all subjects who had spine BMD measurements had them made at baseline.

Using discriminant analysis allows us to examine differences between groups in a number of risk factors simultaneously, rather than having to perform separate comparisons for each risk factor as Black et al. [33]did. This will offer greater power to detect differences between methods than several separate tests. However, it has the drawback that it is not possible to compare as many different methods using this technique as Black et al. did, since each subject needs to be assigned to a subgroup according to which definitions of deformity it satisfies. The number of subgroups to be compared increases exponentially with the number of methods compared, and the number of subjects in each group becomes small, making reliable comparisons impossible. It is, therefore, possible that combining a different criterion for the minimum change with a different criterion for a prevalent deformity may yield even better results than the combination we have presented. However, the 20% height reduction and McCloskey-Kanis criterion for a new prevalent deformity were chosen not only because they are widely used, but also because Black et al. did not find any methods to perform better than them.

The combination method can identify a vertebra as an incident deformity even if it was already deformed at the start of the study. However, it may be necessary to treat such deformities differently from deformities in previously undeformed vertebrae. It has been shown that changes in height between films tend to be greater

in deformed vertebrae [11], possibly because point placement is more difficult. This would lead to a greater false positive rate in previously deformed vertebrae than in undeformed vertebrae. However, since there were only 7 such vertebrae in this study, the problem may be of little practical consequence.

This study could be criticised for using the subject as the unit of analysis, rather than the vertebra. Looking at individual vertebrae with straightforward discriminant analyis would not be valid, due to the fact that vertebrae within an individual are not independent (the risk of a fracture may depend on factors, such as bone quality, which vary between individuals). However, it would have been possible to use a hierarchical model, which allows for these within-subject correlations, to look at the effects of subject-level variables on individual vertebrae. The main reason for not using such models was the difficulty of presenting such models to the target audience, which is to a large extent the clinicians involved in clinical trials of osteoporosis drugs, who would prefer to see simpler statistical models. Some work using hierarchical models to predict incident fractures using both subject-level and vertebral-level predictors has been performed, but it has been difficult to get it published in the osteoporosis literature due to its statistical complexity.

15. Established Incident Models

15.1. Introduction

In this chapter we will look at how well some published morphometric definitions of incident deformity perform. We will consider the three methods: the point prevalence method (using the McCloskey-Kanis definition of a prevalent deformity), the height reduction method proposed by the Study of Osteoporotic Fractures, and the combination method proposed in Chapter 14.

15.2. Methods

The definitions of the three methods to be compared have been given in Chapter 14. The proportion of vertebra classed as incident deformities was calculated for 5 groups of vertebrae:

1. All vertebrae in subjects with no deformities, plus those vertebrae classed as unfractured by the radiologist in those subjects with fractures (prevalent or

incident).

2. Vertebra classed as prevalent fractures in subjects with no incident fractures.

3. Vertebrae classed as prevalent fractures which did not change type in subjects with incident fractures.

4. Vertebra which changed fracture type (or changed from unfractured to fractured) in subjects with incident fractures.

5. Vertebrae in subjects classed as having deformities other than fractures.

It is not certain whether the vertebrae in group 3 had incident fractures or not. It is possible that there was no change in these vertebrae, but it is also possible that the fracture got markedly worse between the two x-rays. This is because the diagnosis of an incident fracture was given at the *subject* level, rather than at the vertebral level. Since we cannot distinguish between these two possibilities given the data available, these vertebrae need to be treated separately.

A similar situation holds for the vertebrae in subjects who had deformities other than fractures. We again know that there is a deformity somewhere in the spine, but not precisely where. However, *none* of these vertebrae can be incident fractures, and so any classed as such by any of our methods are false positives.

15.3. Results

The number of vertebrae classed as incident deformities by each of the three methods in each of the five groups of vertebrae are given in Table 15.1

Vertebral Group	Point Prevalence	Height Reduction	Combination
1	169/69967 = 0.24%	49/69967 = 0.07%	17/69967 = 0.02%
2	53/ 854 = 6.2 %	17/ 854 = 2.0%	16/ 854 = 1.9%
3	14/ 163 = 8.6%	30/ 163 = 18.4%	30/ 163 = 18.4%
4	241/ 295 = 81.7%	227/ 295 = 77.0%	224/ 295 = 75.9%
5	55/ 5326 = 1.0%	12/ 5326 = 0.23%	8/ 5326 = 0.15%

Table 15.1.: Numbers of vertebrae classed as incident deformities by existing methods.

The most sensitive method is the point prevalence method, detecting 241 of the 295 genuine incident fractures in this population, i.e. of the vertebrae in group 4. However, this method also classified 169 vertebrae in subjects with no deformities as incident fractures, compared to only 49 using the height reduction method and 17 using the combination method. This means that it produces over 3 times the false positive rate of the height reduction method and nearly ten times the false positive rate of the combination method. It should be noted that the false positive rates in Table 15.1 are per vertebra, and 13 vertebrae are examined in each individual. Thus, the false positive rates per subject will be considerably higher (around 3%, 1% and 0.3% for subjects with no fractures for the point prevalence, height reduction and combination methods respectively).

One odd result in Table 15.1 is that for group 3, in which only 14 vertebrae were fractures using the point prevalence method but 30 were using the combination method. This can be explained as follows: the combination method makes no

assumption about the state of the vertebra on the first film, whilst the point prevalence method insists that it was not classed as a deformity on this film. These vertebrae are those that were classified as prevalent fracures on the first film by the radiologist, and 135 / 169 were classed as prevalent deformities, and therefore could not satisfy the point prevalence criterion. However, if they showed a further reduction of 20%, they could satisfy the combination criterion.

15.4. Discussion

These results are very similar to those seen in the previous chapter. This is hardly surprising, since the same methods were used on a subset of the data. Again, we conclude that the point prevalence method lacks specificity. The combination method is slightly less sensitive but considerably more specific than the height reduction method, and is therefore preferable in situations where specificity is particularly important (i.e. where the incidence of genuine fractures is low).

16. Identifying Incident Deformities Using Polynomial Models

16.1. Introduction

In this chapter, we will be looking at how we can use the predicted heights from the polynomial models devised in Chapter 6 to identify incident vertebral deformities. The heights measured on both the first and second round x-rays will be compared to their predicted values from the polynomial model, and the residuals from both of these model fittings used to identify incident fractures.

16.2. Methods

The methods outlined in Chapter 6 were used to predict vertebral heights. Only 2 models were used, one for men and one for women. The 70 subjects used to generate these models were randomly selected from all centres. Two separate magnification

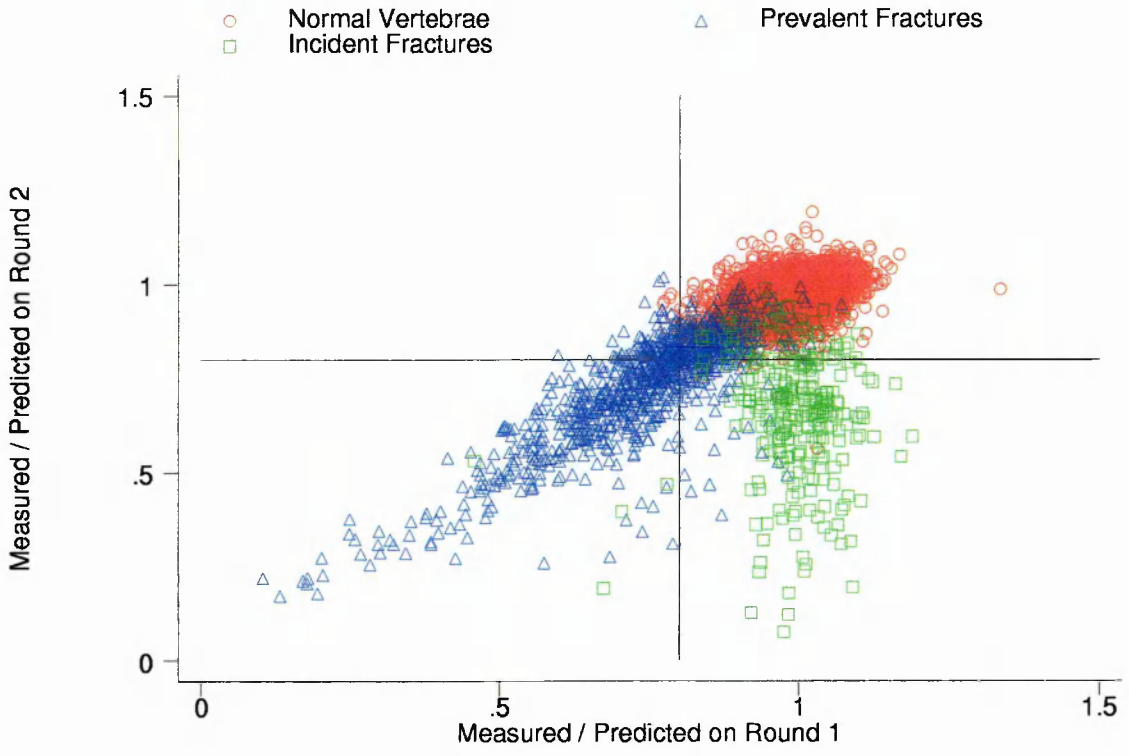


Figure 16.1.: Observed height / predicted height from polynomial models on first and second round x-rays

factors were calculated for each subject, one for the first round measurements and one for the second round measurements.

Having obtained predicted heights for each subject on both rounds, any height that is less than 80% of its predicted height is considered to be a deformity.

16.3. Results

16.3.1. Examination of Residuals

Figure 16.1 show how the observed height/predicted heights from the polynomial models on the first and second round x-rays can be used to differentiate between undeformed vertebrae, prevalent fractures and incident fractures. The undeformed vertebrae (in red) tend not to have reductions in height on either round of x-rays. Prevalent deformities (in blue) tend to have reductions in height on both x-rays, and incident deformities (in green) tend to have reductions in height on the second round but not the first. It is possible for a vertebra to be both a prevalent and incident deformity, if it had lost height at the time of the first x-ray and then lost more height between the two x-rays. Such vertebrae appear between the solid area of blue and the solid area of green.

Given the large number of points in Figure 16.1, it is not possible to see every point. Therefore the 3 types of vertebrae have been plotted separately in Figures 16.2, 16.3 and 16.4.

Figure 16.2 shows that the bulk of heights in vertebrae judged to be undeformed by the radiologist are more than 80% of their expected heights on both x-rays. However, there are a small number of vertebrae that are less than 2/3 of their expected heights on one or both films.

Figure 16.3 shows that the majority of vertebra classed as prevalent deformities lie close to the line $y = x$, i.e. the measured heights are less than expected, but are similar on round 1 and round 2. However, there is a small group to the right of

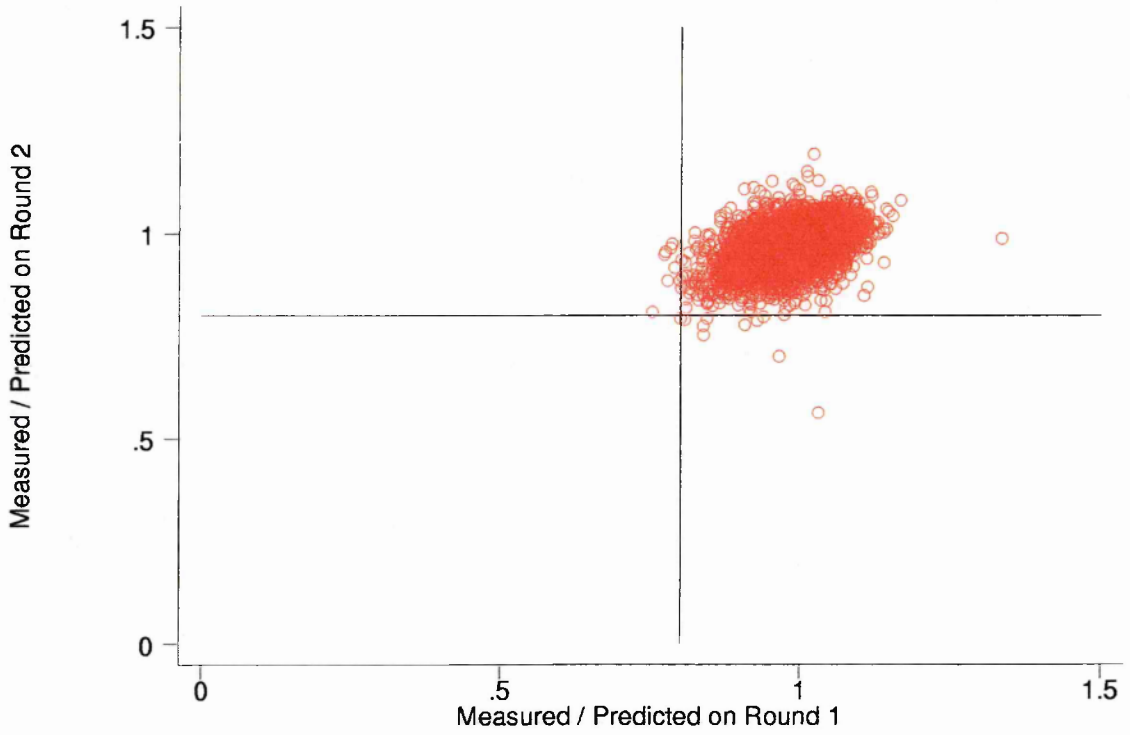


Figure 16.2.: Observed height / predicted height from polynomial models on first and second round x-rays in undeformed vertebrae

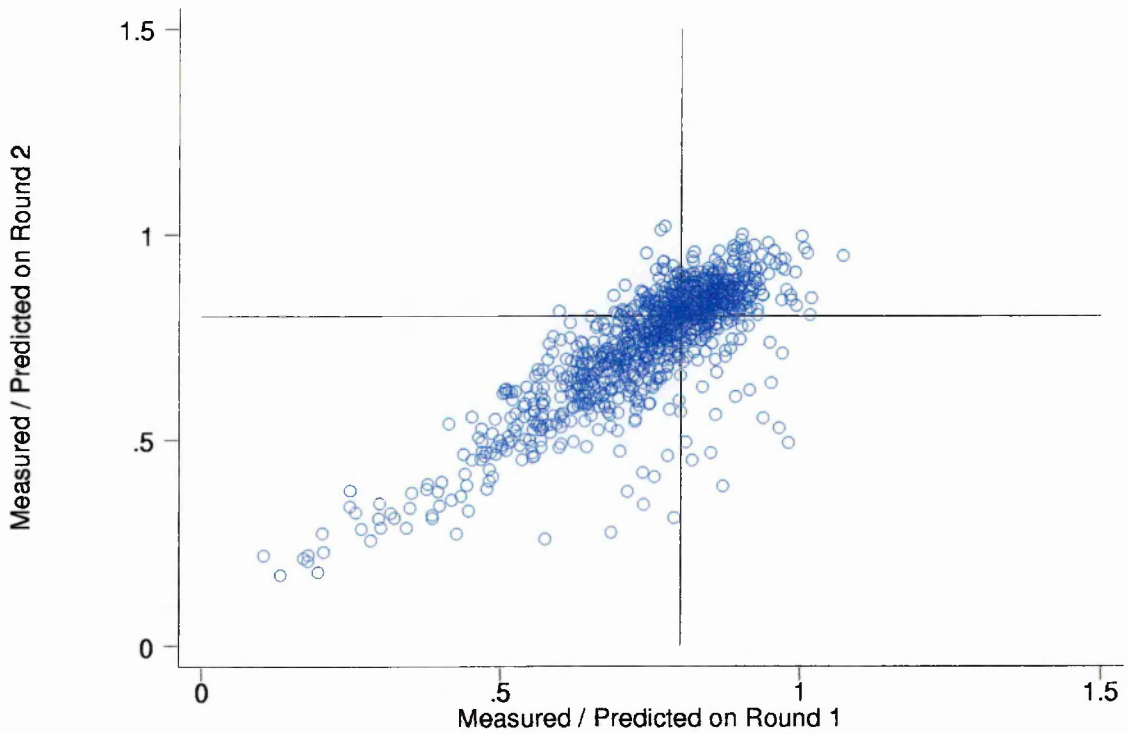


Figure 16.3.: Observed height / predicted height from polynomial models on first and second round x-rays in prevalent fractures

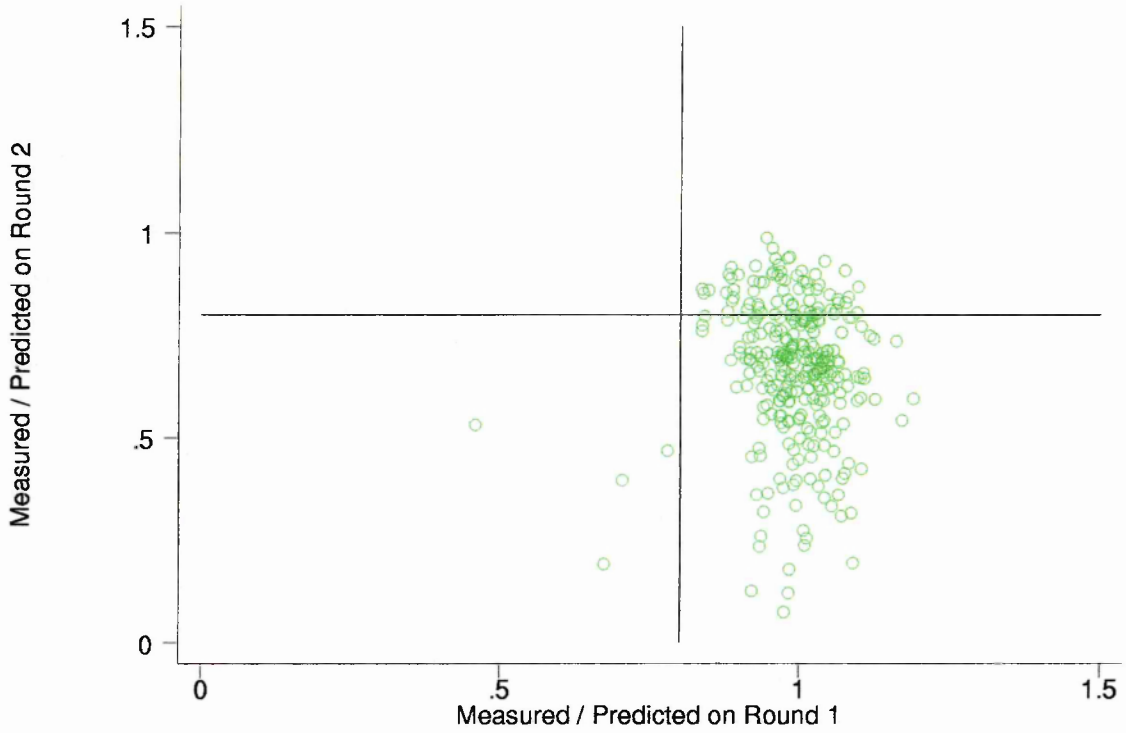


Figure 16.4.: Observed height / predicted height from polynomial models on first and second round x-rays in incident fractures

this line, which showed some reduction in height on round 1 but a greater reduction on round 2. In addition, whilst many of the vertebrae show a loss of height on both rounds (i.e. they are below and to the left of the reference lines), there are a considerable number of fractures which do not show a loss of height.

Figure 16.4 shows the vertebrae classed as incident deformities. The vast majority of these show no height reduction on round 1, but some reduction on round 2. However, there are a small number which show some reduction on round 1, and either the same or greater reduction on round 2.

16.4. Identification of Incident Fractures

16.4.1. Point Prevalence Method

Points to the left of the vertical line in Figure 16.1 are vertebrae that are classed as prevalent deformities on the first x-ray, whilst points below the horizontal line are classed as deformities on the second x-ray (using the polynomial method). Thus, if we use the point prevalence definition of incident deformities, points in the lower right quadrant of this figure represent incident deformities. The number of normal vertebrae, prevalent fractures and incident fractures classed as incident deformities using this method is given in Table 16.1.

This method has many fewer positives in group 1 than the McCloskey-Kanis point prevalence method shown in Table 15.1, but rather more in groups 2 & 3. On the other hand, it has slightly fewer true positives in group 4. Compared to the height reduction and combination methods shown in Table 15.1, this method

Vertebral Group	Deformities
1	78/69967 = 0.11%
2	113/ 854 = 13.2%
3	31/ 163 = 19.0%
4	224/ 295 = 75.9%
5	56/ 5326 = 1.05%

Table 16.1.: Incident deformities defined by polynomial models: point prevalence method

has many more false positives in groups 1, 2, 3 and 5, and only slightly more true positives in group 4. It is therefore less good than these alternative methods.

16.4.2. Combination Method

We saw in Chapters 14 and 15 that the point prevalence method (using the McCloskey-Kanis method of defining prevalent deformities) led to a large number of false positives, and a combination of a prevalent deformity criterion and a change in height criterion performed better. A similar approach can be used with this polynomial model.

Figure 16.5 is similar to Figure 16.1, but instead of the relative height reduction on the first round, in this case the relative change in height between the two films is plotted on the x-axis. Thus heights that have shown a marked reduction in height between the two films are on the left of this plot, and heights that are significantly less than their expected values are towards the bottom of the plot. The

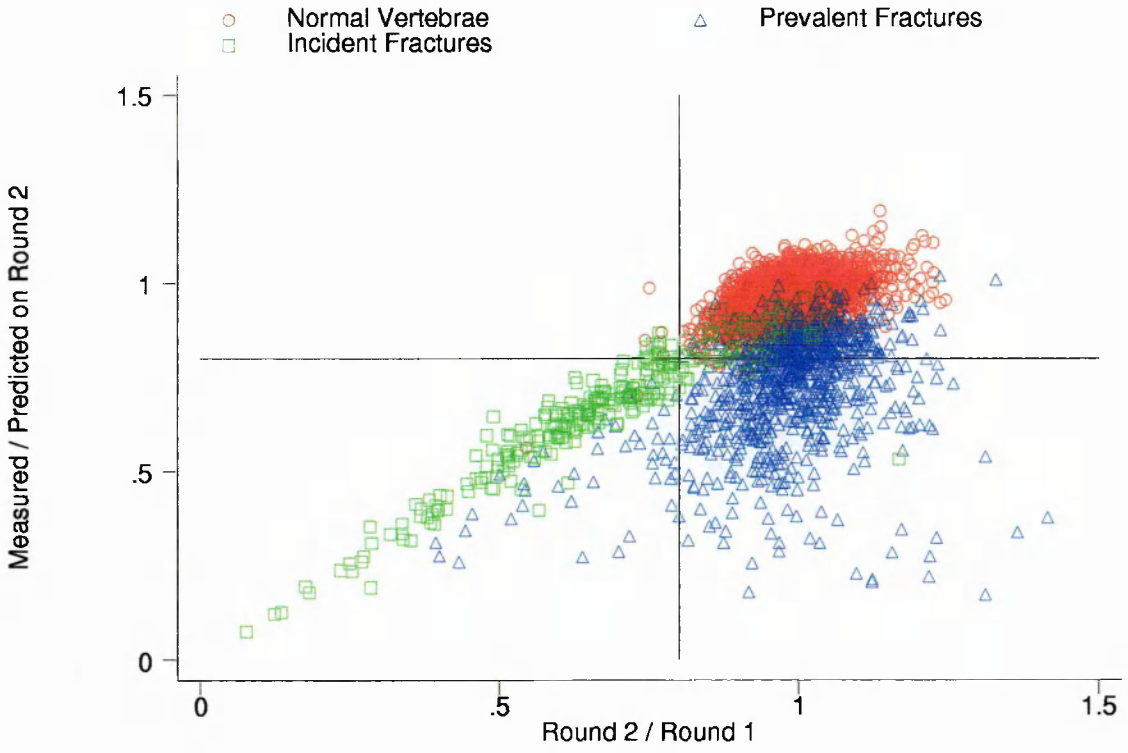


Figure 16.5.: Plot of observed height / predicted height against relative change in height between first and second round x-rays using polynomial models

points plotted in the lower left quadrant are therefore those that can be classified as incident deformities. The number of normal vertebrae, prevalent fractures and incident fractures classed as incident deformities using this method is given in Table 16.2.

These results are extremely similar to those for the combination method using the McCloskey-Kanis algorithm shown in Table 15.1. However, there are 4 more false positives and 10 more false negatives using this method, so this method cannot be claimed to be an improvement.

Vertebral Group	Deformities
1	18/69967 = 0.03%
2	17/ 854 = 2.0%
3	30/ 163 = 18.4%
4	214/ 295 = 72.5%
5	10/ 5326 = 0.19%

Table 16.2.: Incident deformities defined by polynomial models: combination method

16.5. Discussion

Using a polynomial model to identify prevalent fractures rather than the McCloskey-Kanis algorithm led to very little difference using either the point prevalence or combination approaches.

The sensitivity of all of the morphometric methods are probably being underestimated, since there is one individual who was classed as being an incident case but no vertebrae were classed as fractures on the second round. It is therefore likely that the individual has at least one fracture, but that it was not recorded by the radiologist. According to all of the morphometric methods, there were three incident fractures in this subject. It is likely that at least one of these vertebrae had a genuine fracture, for the subject to be classed as an incident case, and hence will be counted as a false positive when in fact it is a true positive. It is possible that that the classification of the subject was incorrect, but given that the radiologist and all

three morphometric methods agree that the subject suffered an incident deformity, that is unlikely.

There are other vertebrae in which one or more heights are very much less than expected, and yet the vertebra is classed as normal by the radiologist. These may also be fractures that the radiologist did not record, but they may also be incorrectly recorded heights. In cases where the heights are very low on both x-rays, it is more likely that there is a genuine fracture, since making the same error on both films is unlikely. However, if the low height in only on the second film, the error could be either in the height or in not identifying a fracture. So we know that there are some fractures that have not been recorded, but we cannot know how many.

On the other hand, there are a large number of vertebrae that are classed as fractures by the radiologist that show no appreciable reduction in height. In the main, these are classed as wedge deformities. In these cases, it is possible that there is a qualitative feature of the x-ray apparent to the radiologist that is not detected by measuring the vertebral heights. Hence these cases will not be detected by morphometry. We cannot conclude that there are any false positives in the clinical classification.

17. Using the Imputation Method to Define Incident Deformities

17.1. Introduction

We have seen that combining a measure of change with a definition of prevalent deformity is the best method of defining incident fractures. In Part II, we saw that the imputation method of defining prevalent deformities performed better than any of the others that we considered. It therefore seemed reasonable to test whether using this definition of deformity, together with a measure of change, provided a better definition of incident deformity than those we have seen so far.

17.2. Methods

The methods outlined in Chapter 10 were used to define prevalent deformities, using the second round height measurements. Then the relative loss of height between

the first and second round x-rays was calculated for each measured height. Any vertebra which had lost 20% of its height between the two x-rays, and was classed as a prevalent deformity on the second film, was considered an incident deformity by this method.

This method was compared to the method presented previously in Chapter 14, which used the McCloskey-Kanis method for the "shape" part of the incident deformity definition. Again, the radiologist's reading was used as a gold standard. Only vertebrae that could be assessed by both the imputation method and the McCloskey-Kanis method were included in the analysis. Men from 2 centres had to be excluded from the analysis completely, since these centres did not provide enough male subjects to define the imputation model.

17.3. Results

Of 72,272 vertebrae that could be assessed, 328 were classed as incident fractures by the radiologist. The imputation method defined slightly fewer of these to be fractures than the original combination method (245 vs 250), but the difference was not statistically significant using McNemar's test. The imputation method also defined as fractures fewer of the vertebrae declared as not fractured by the radiologist (46 vs 53), but again the difference was not statistically significant using McNemar's test.

Many of the vertebrae classed as incident fractures by the morphometric methods, but not by the radiologist, were classed as prevalent fractures by the radiolo-

gist. In fact, the two methods only differed in the classification of vertebrae classed as undeformed by the radiologist. Details are given in Table 17.1.

Shape	Imputation Method	McCloskey-Kanis Method
Normal	19	26
Wedge	9	9
Concavity	13	13
Biconcavity	2	2
Crush	3	3

Table 17.1.: Clinical shapes of false positive morphometric vertebral deformities

17.4. Discussion

Using the imputation method rather than the McCloskey-Kanis method for the "shape" component of a combination method of defining incident fractures did not provide a significant improvement. This was at first surprising, since the imputation method had only half of the false positives of the McCloskey-Kanis method when considering prevalent deformities.

However, the methods need only be applied to vertebrae that have shown a reduction in height of at least 20% in at least one vertebra. Therefore, there are a far smaller number of vertebrae being tested as possible prevalent deformities (345 vs 79,272).

In addition, many of the false positives are in fact prevalent, rather than incident,

fractures according to the radiologist. It has been shown that vertebral height measurements are less reproducible in deformed vertebrae, and it may be that the increased measurement error is responsible for the apparently large reductions in height between the two films. However, it is also possible that there has been some progression in the deformity between the two films, which has not been noticed by the radiologist. A reduction of 20% would be noticed, but a slight reduction combined with increased measurement imprecision may explain the increased false positive rate.

18. Comparison of Incident Deformity Models

We saw in Chapter 14 that combining two criteria, one relating to the shape of the vertebra on the second film and one relating to the height lost since the first film, gives better agreement with the radiologists opinion than either criterion individually. The choice of method then becomes a choice for each of these criteria.

A height loss of 20% is currently widely used. Part of the reason for this is that it has been accepted by the American Food and Drugs Administration as a suitable definition for vertebral fracture in clinical trials. Therefore, it seems reasonable to retain this threshold, and compare different choices for the "shape" criterion.

It has been suggested that this method is simply equivalent to choosing a more stringent threshold for the height loss criterion. This is not the case: the shape of the vertebra on the second film is not directly related to the amount of height lost between the two films.

In the analysis in Chapter 17, there were 345 vertebrae which lost 20% of their height or more, of which 303 also satisfied the McCloskey-Kanis algorithm. Of the

42 vertebrae excluded by the second condition, only 2 were classed as fractures by the radiologist. To eliminate 42 vertebrae by making the threshold more stringent would require a threshold of 21.92%, but of the 42 vertebrae eliminated in this way, 11 were considered to be fractures by the radiologist. There using two different criteria is having less of a negative impact on sensitivity than using a more stringent threshold.

There is also the advantage with this method that a vertebra classed as an incident fracture must be classed as a prevalent fracture. This makes comparisons between incident and prevalent fractures easier.

There is still a choice to be made about which particular criterion to use for the shape part of the definition. In Part II, we saw that the McCloskey-Kanis and polynomial models gave very similar results when it came to identifying prevalent deformities, whilst the imputation method was somewhat better. However, it made little difference which method was chosen when identifying incident deformities. This may be because only vertebrae that have shown a change of 20% or more between the two films are considered as potential deformities, rather than all vertebrae. Not only does this much smaller number of vertebrae tested reduce the number of false positives, but the fact that there has been a marked reduction in height increases the probability that the vertebra is an unusual shape. Therefore, the choice of shape criterion is less critical when defining incident deformities than when defining prevalent deformities. It therefore makes sense to use the same criterion for both incident and prevalent deformities.

Morphometric methods appear to work better at identifying incident deformities

than they do at identifying prevalent deformities. This is due in part to differences in shape between prevalent and incident fractures. This is shown in Table 18.1: over half of the prevalent deformities are wedges, compared to only one quarter of incident deformities. Since these deformities are more difficult to detect morphometrically, it will be harder to detect prevalent rather than incident deformities.

Shape	Prevalent		Incident	
Wedge	558	(54%)	86	(26%)
Concavity	403	(39%)	198	(60%)
Biconcavity	65	(6%)	34	(10%)
Crush	14	(1%)	10	(3%)
Total	1040		328	

Table 18.1.: Shapes of incident and prevalent fractures

Incident deformities also tend to be larger than prevalent deformities when compared using morphometric criteria. For example, the median deformity severity using the McCloskey-Kanis algorithm was 4.4 standard deviations in those vertebra judged to be prevalent deformities by the radiologist, compared to 6.1 standard deviations for the incident deformities. This difference was highly statistically significant using the Wilcoxon signed rank sum test. This also makes it easier to detect them morphometrically.

Another advantage of considering incident deformities is that many of the processes that can lead to deformity work more slowly than fractures. For example, degenerative change due to osteoarthritis occurs gradually, and congenital defor-

mities do not change at all. These types of deformity are therefore unlikely to be identified as incident fractures.

The main drawback of using incident fractures as an endpoint is the fact that they are so much less common than prevalent deformities. This is inevitable, since incident deformities have to occur within a predetermined time period, whilst prevalent deformities can have occured at any time during the subject's life. Thus a larger study will be required to obtain the same number of fractures.

Bibliography

[1] L. M. Hurxthal. Measurements of vertebral heights. *American Journal of Roentgenology*, 103:635–644, 1968.

[2] E. Barnett and B. E. C. Nordin. The radiological diagnosis of osteoporosis; a new approach. *Clinical Radiology*, II:166–174, 1960.

[3] L. Ferrar, N. A. Barrington, G. Jiang, and R. Eastell. Vertebral morphometry by morphometric radiography (MR) and morphometric x-ray absorptiometry(MXA) using the same reference population. *Journal of Bone and Mineral Research*, 12(S1):s178, August 1997. Abstract.

[4] P. D. Ross, J. W. Davis, R. S. Epstein, and R. D. Wasnich. Pre-existing fracture and bone mass predict vertebral fracture incidence in women. *Annals of Internal Medicine*, 114(11):919–923, June 1991.

[5] R. Smith-Bindman, S. R. Cummings, P. Steiger, and H. K. Genant. A comparison of morphometric definitions of vertebral fracture. *Journal of Bone and Mineral Research*, 6(1):25–34, 1991.

[6] H. W. Minne, G. Leidig, C. Wüster, L. Siromachkostov, G. Baldauf, R. Bickel, P. Sauer, M. Lojen, and R. Ziegler. A newly developed spine deformity index (SDI) to quantitate vertebral crush fractures in patients with osteoporosis. *Bone and Mineral*, 3:335–349, 1988.

[7] J. C. Gallagher, L. R. Hedlund, S. Stoner, and C. Meeger. Vertebral morphometry: Normative data. *Bone and Mineral*, 4:189–196, 1988.

[8] L. J. Melton III, S. H. Kan, M. A. Frye, H. W. Wahner, W. M. O'Fallon, and B. L. Riggs. Epidemiology of vertebral fractures in women. *American Journal of Epidemiology*, 129(5):1000–1010, 1989.

[9] R. Eastell, S. L. Cedel, H. W. Wahner, B. L. Riggs, and L. J. Melton III. Classification of vertebral fractures. *Journal of Bone and Mineral Research*, 6(3):207–215, 1991.

[10] D. M. Black, L. Palermo, M. C. Nevitt, H. K. Genant, R. Epstein, R. San Valentin, and S. R. Cummings. Comparison of methods for defining prevalent vertebral deformities: The study of osteoporotic fractures. *Journal of Bone and Mineral Research*, 10(6):890–902, 1995.

[11] E. McCloskey, T. D. Spector, K. S. Eyres, E. D. Fern, N. O'Rourke, S. Wasikaran, and J. A. Kanis. The assessment of vertebral deformity: a method for use in population studies and clinical trials. *Osteoporosis International*, 3(3):138–147, 1993.

[12] M. Lunt, D. Felsenberg, J. Reeve, L. I. Benevolenskaya, J. Cannata, J. Deque-

ker, C. Dodenhof, J. A. Falch, P. Masaryk, H. A. P. Pols, G. Poor, D. M. Reid, C. Scheidt-Nave, K. Weber, J. Varlow, J. A. Kanis, T. W. O'Neill, and A. J. Silman. Bone density variation and its effects on risk of vertebral deformity in men and women studied in 13 European centres: the EVOS study. *Journal of Bone and Mineral Research*, 12(11):1883–1894, November 1997.

[13] National Osteoporosis Foundation Working Group on Vertebral Fractures. Assessing vertebral fractures. *Journal of Bone and Mineral Research*, 10(4):518–523, Apr 1995.

[14] J. A. Cauley, P. A. Murphy, T. J. Riley, and A. M. Buhari. Effects of fluoridated drinking water on bone mass and fractures: The Study of Osteoporotic Fractures. *Journal of Bone and Mineral Research*, 10(7):1076–1086, Jul 1995.

[15] M. Lunt, A. Ismail, D. Felsenberg, C. Cooper, J. Kanis, J. Reeve, A. Silman, T. O'Neill, and the EPOS Study Group. Defining incident vertebral deformities in population studies: A comparison of morphometric criteria. *Osteoporosis International*, 2002. *(In Press)* .

[16] D. M. Black, S. R. Cummings, K. Stone, E. Hudes, L. Palermo, and P. Steiger. A new approach to defining normal vertebral dimensions. *Journal of Bone and Mineral Research*, 6(8):883–892, 1991.

[17] D. M. Rocke and D. L. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91(435):1047–1061, September 1996.

[18] The European Prospective Osteoporosis (EPOS) Study Group. Incidence of vertebral fracture in Europe: Results from the European Prospective Osteoporosis Study (EPOS). *Journal of Bone and Mineral Research*, 17(4):716–724, 2002.

[19] M. Mittlbock and M. Schemper. Explained variation for logistic regression. *Statistics in Medicine*, 15(19):1987–1997, 1996.

[20] G. G. Judge, W. E. Griffiths, R. Hill, H. Lütkepohl, and T.-C. Lee. *The Theory and Practice of Econometrics.* John Wiley & Sons, New York, 2 edition, 1985.

[21] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845, 1988.

[22] F. Mosteller and J. W. Tukey. *Data Analysis and Regression.* Addison-Wesley, Reading, Mass., 1977.

[23] A. S. Hadi. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society, Series B.*, 54(3):761–771, 1992.

[24] A. S. Hadi. A modification of a method for the detection of outliers in multivariate samples. *Journal of the Royal Statistical Society, Series B.*, 56(2):393–396, 1994.

[25] P. J. Rousseeuw and B. C. van Zomeren. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411):633–639, September 1990.

[26] K. I. Penny. Appropriate critical values when testing for a single multivariate outlier by using the mahalanobis distance. *Applied Statistics*, 45(1):73–81, 1996.

[27] D. M. Rocke. Robustness properties of s-estimators of multivariate location and shape in high dimension. *The Annals of Statistics*, 24:1327–1345, 1996.

[28] W. J. Krzanowski. *Principles of Multivariate Analysis*, chapter 8.1. Oxford University Press, Oxford, 1988.

[29] P. Royston and D. G. Altman. Approximating statistical functions by using fractional polynomial regression. *Statistician*, 46(3):411–422, 1997.

[30] L. J. Melton, A. W. Lane, C. Cooper, R. Eastell, W. M. O'Fallon, and B. L. Riggs. Prevalence and incidence of vertebral deformities. *Osteoporosis International*, 3(3):113–119, 1993.

[31] D. M. Black, S. R. Cummings, D. B. Karpf, J. A. Cauley, D. E. Thompson, M. C. Nevitt, D. C. Bauer, H. K. Genant, W. L. Haskell, R. Marcus, S. M. Ott, J. C. Torner, S. A. Quandt, T. F. Reiss, K. E. Ensrud, and the Fracture Intervention Trial Research Group. Randomised trial of the effect of alendronate on risk of fracture in women with existing vertebral fractures. *Lancet*, 348:1535–41, Dec 1996.

[32] B. L. Riggs, E. Seeman, S. F. Hodgson, T. D. R., and W. M. O'Fallon. Effect of the fluoride calcium regimen on vertebral fracture occurence in postmenopausal osteoporosis. *New England Journal of Medicine*, 306:446–450, 1982.

[33] D. M. Black, L. Palermo, M. C. Nevitt, H. K. Genant, L. Christensen, and S. R. Cummings. Defining incident vertebral deformity: A prospective comparison of several approaches. *Journal of Bone and Mineral Research*, 14(1):90–101, 1999.

[34] J. Dequeker, J. Pearson, J. Reeve, M. Henley, J. Bright, D. Felsenberg, W. Kalender, A.-M. Laval-Jeantet, P. Ruegsegger, J. Adams, M. Diaz Curiel, M. Fischer, F. Galan, P. Geusens, L. Hyldstrup, P. Jaeger, P. Kotzki, H. Kroger, P. Lips, A. Mitchell, O. Louis, R. Perez Cano, H. Pols, D. M. Reid, C. Ribot, P. Schneider, and M. Lunt. Dual X-ray Absorptiometry – cross-calibration and normative reference ranges for the spine: results of a European Community Concerted Action. *Bone*, 17(3):247–254, 1995.

[35] J. Pearson, J. Dequeker, J. Reeve, D. Felsenberg, M. Henley, J. Bright, M. Lunt, J. Adams, F. Galan, P. Geusens, P. Jaeger, H. Kroger, P. Lips, A. Mitchell, R. Perez Cano, H. Pols, D. M. Reid, C. Ribot, P. Schneider, A.-M. Laval-Jeantet, P. Ruegsegger, and W. Kalender. Dual X-ray absorptiometry of the proximal femur: normal European values standardised with the European Spine Phantom. *Journal of Bone and Mineral Research*, 10:315–324, 1995.

[36] J. Pearson, J. Dequeker, M. Henley, J. Bright, J. Reeve, W. Kalender, A. M. Laval Jeantet, P. Ruegsegger, D. Felsenberg, J. Adams, and others. European semi-anthropomorphic spine phantom for the calibration of bone densitometers: assessment of precision, stability and accuracy. The European Quantitation of Osteoporosis Study Group. *Osteoporosis International*, 5(3):174–84, May 1995.

[37] M. Lunt, D. Felsenberg, J. Adams, L. Benevolenskaya, J. Cannata, J. Dequeker, C. Dodenhof, J. Falch, O. Johnell, K.-T. Khaw, P. Masaryk, H. Pols, G. Poor, D. Reid, C. Scheidt-Nave, K. Weber, A. Silman, and J. Reeve. Population-based geographic variations in DXA bone density in Europe: the EVOS study. *Osteoporosis International*, 7(3):175–189, 1997.

[38] P. D. Ross, Y. K. Yhee, Y.-F. He, J. W. Davis, C. Kamimoto, R. S. Epstein, and R. D. Wasnich. A new method for vertebral fracture diagnosis. *Journal of Bone and Mineral Research*, 8(2):167–174, 1993.

[39] H. K. Genant, M. Jergas, L. Palermo, M. Nevitt, R. S. Valentin, D. Black, and S. R. Cummings. Comparison of semiquantitative visual and quantitative morphometric assessment of prevalent and incident vertebral fractures in osteoporosis: The Study of Osteoporotic Fractures Research Group. *Journal of Bone and Mineral Research*, 11(7):984–996, Jul 1996.

[40] B. Flury and H. Riedwyl. *Multivariate Statistics, A Practical Approach*, chapter 7. Chapman and Hall, London, New York, 1988.

[41] A. A. Ismail, J. D. Finn, M. Lunt, C. Cooper, D. Felsenberg, O. Johnell, J. Reeve, A. J. Silman, T. W. O'Neill, and the EPOS Study Group. Lifestyle factors have only a modest influence in predicting incident vertebral deformity: Results from the European Prospective Osteoporosis Study (EPOS). *Osteoporosis International*, 11(Supp. 2):S96, 2000. abstract.